CS772: Deep Learning for Natural Language Processing (DL-NLP)

Neural POS Tagging, Neural LM

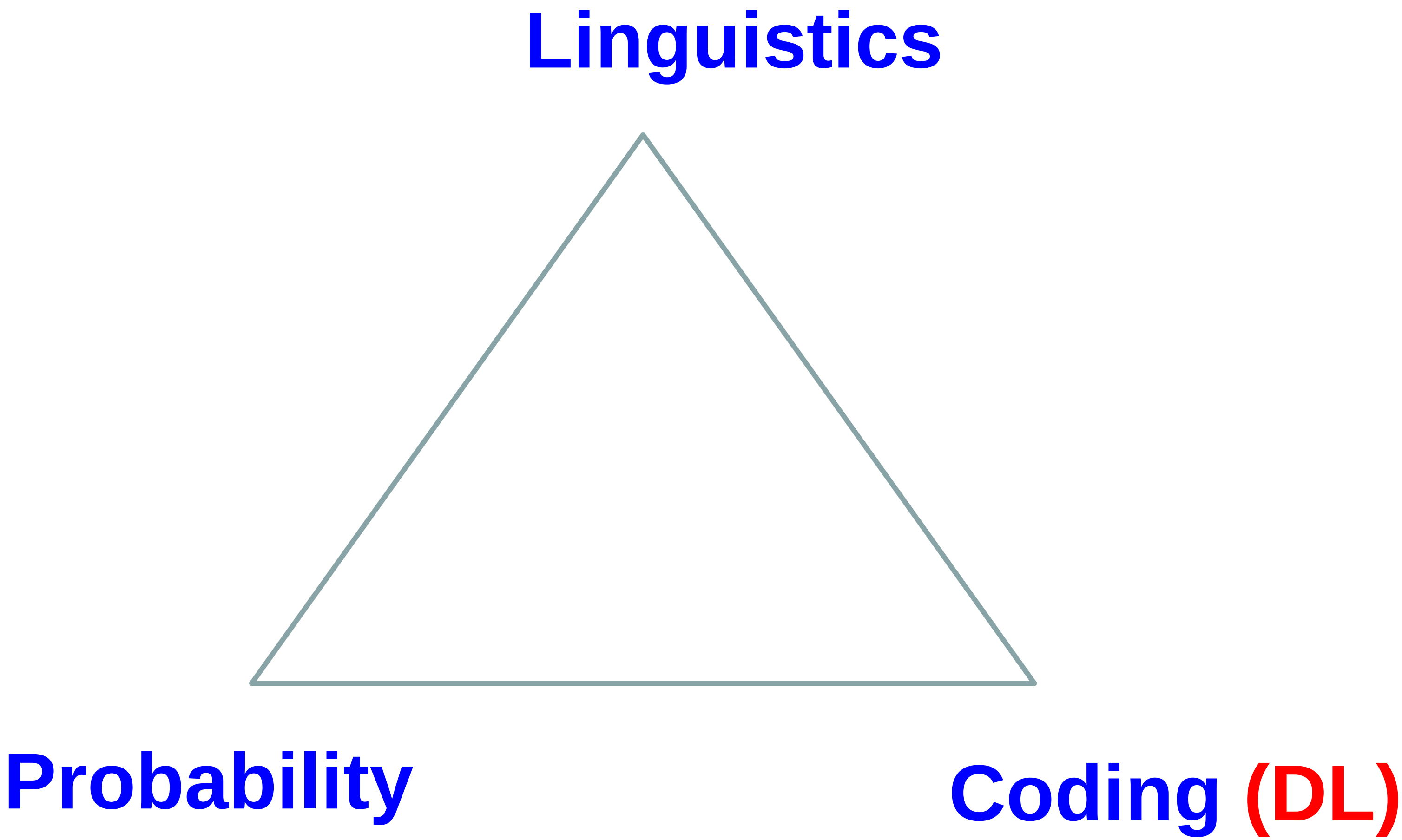
Pushpak Bhattacharyya
Computer Science and Engineering Department
IIT Bombay

Week 2 of 10th Jan, 2022

Task *vs.* Technique Matrix

Task (row) vs. Technique (col) Matrix	Rules Based/Knowledge-Based	Classical ML				Deep Learning		
		Perceptron	Logistic Regression	SVM			RNN-LSTM	CNN
Morphology								
POS		Graphical Models (HMM, MEMM, CRF)	Dense FF with BP and softmax	Graphical Models (HMM, MEMM, CRF)	Dense FF with BP and softmax			
Chunking								
Parsing								
NER, MWE								
Coref								
WSD								
Machine Translation								
Semantic Role Labeling								
Sentiment								

The Trinity of NLP



3 Generations of NLP

- Rule based NLP is also called Model Driven NLP
- Statistical ML based NLP (*Hidden Markov Model, Support Vector Machine*)
- Neural (Deep Learning) based NLP

Illustration with POS tagging

DL-POS

POS tagging problem statement

- Input: sequence of words W
- Output: sequence of tags T

- E.g.
- Input: *I love India*
- Output: PRP VB NNP

Training Data Example: A dialogue text POS tagged from Treebank

[SpeakerB1/SYM]

./.

So/UH how/WRB many/JJ ,/, um/UH ,/,

[credit/NN cards/NNS]

do/VBP

[you/PRP]

have/VB ?/.

[SpeakerA2/SYM]

./.

[Um/UH]

,/,

[I/PRP]

think/VBP

[I/PRP]

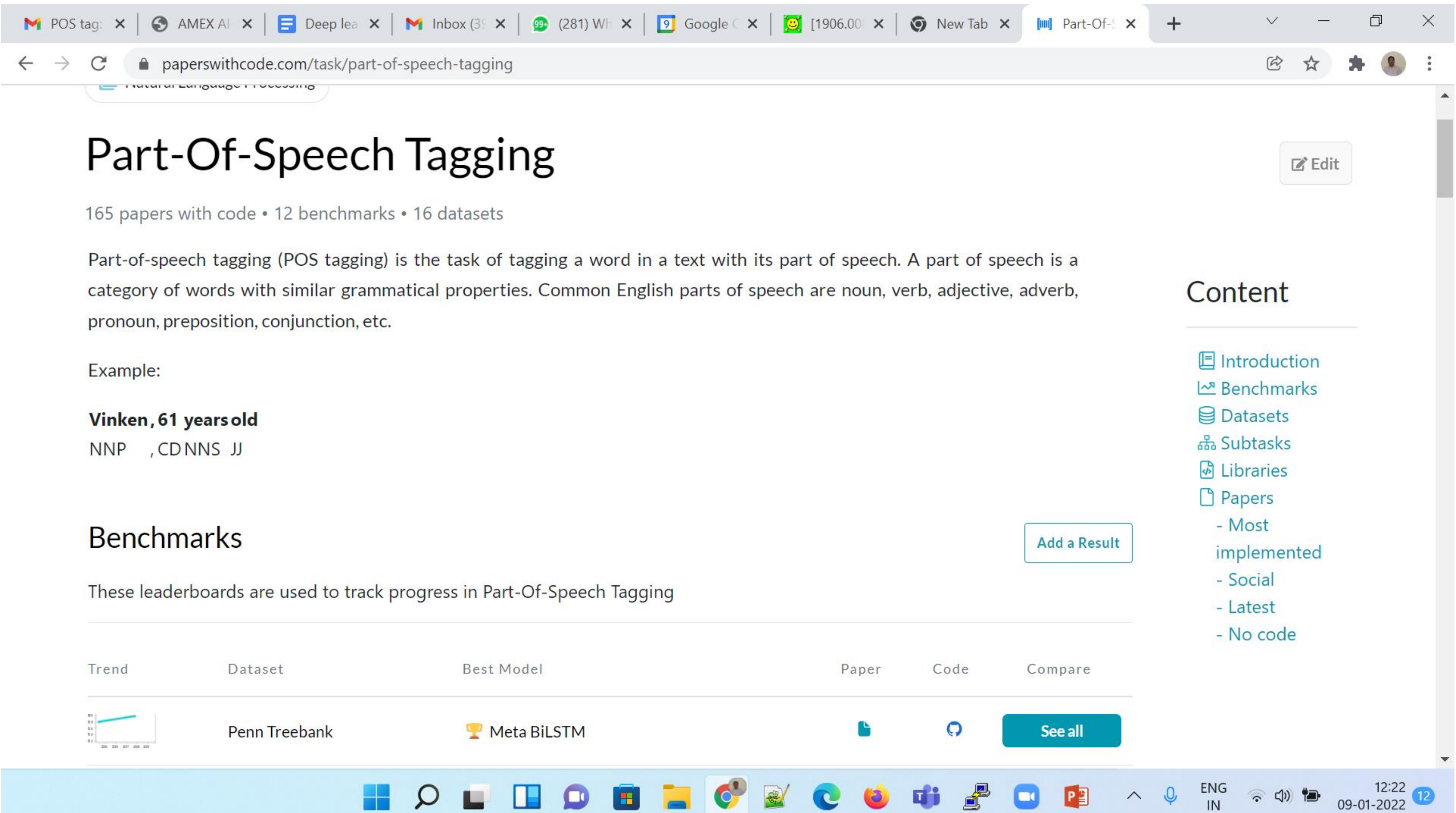
'm/VBP down/IN to/IN

[one/CD]

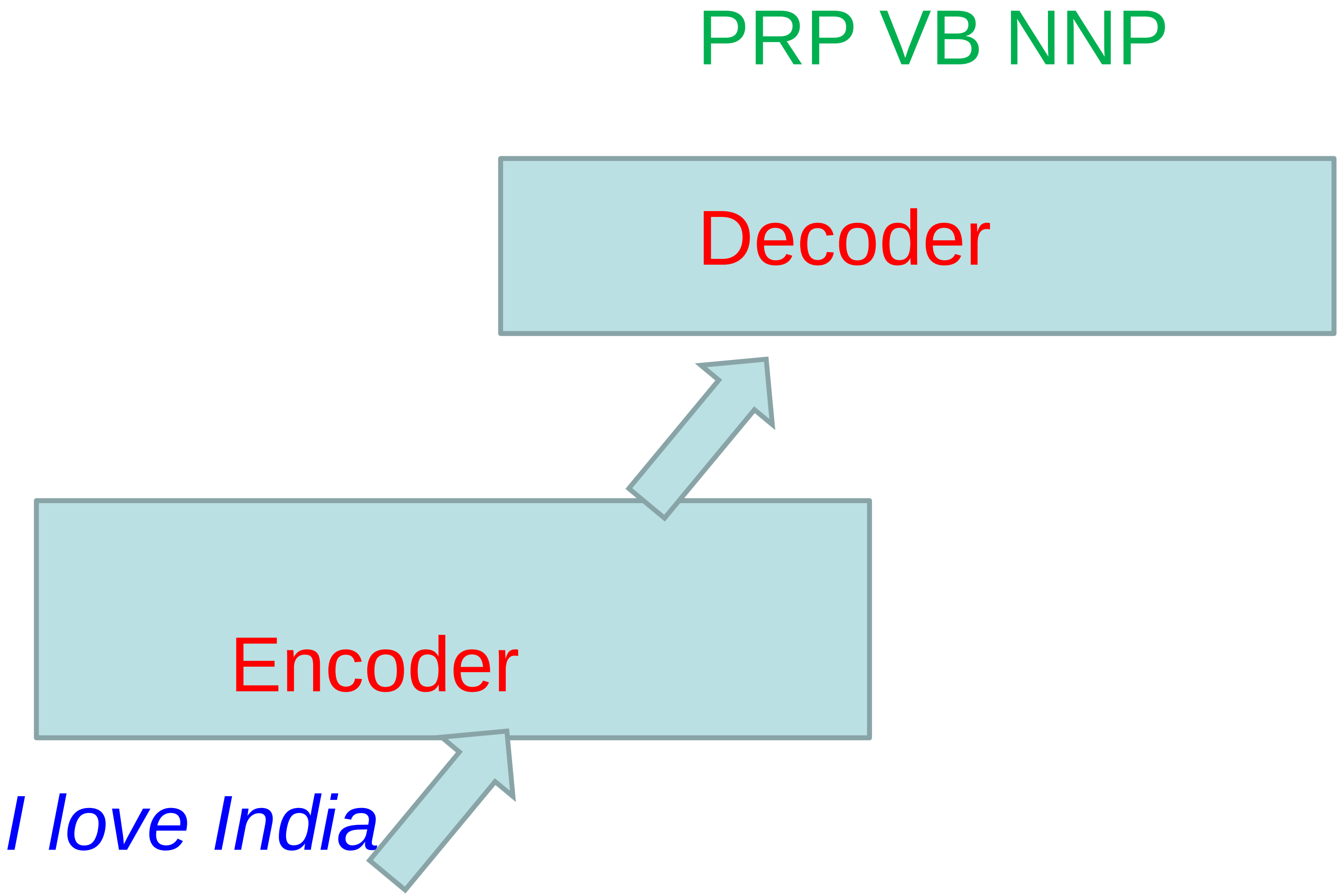
./.

https://catalog.ldc.upenn.edu/desc/addenda/LDC99T42.pos.txt

POS tagging code dataset etc.: paperwithcode.com



DL based POS Tagging



Penn POS TAG Set

1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjun
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative

Penn POS TAG Set (cntd)

22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Minimize Cross Entropy Loss= MLE

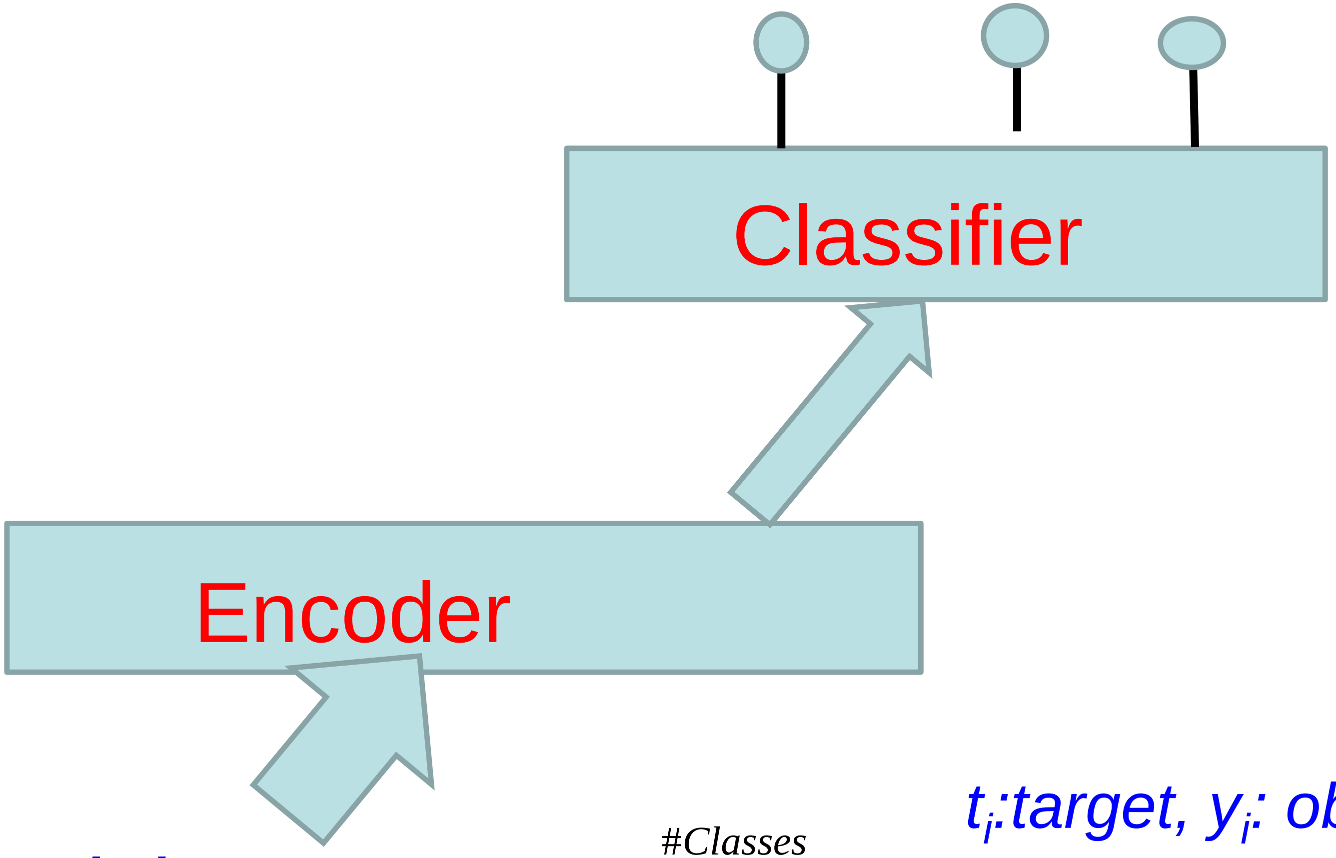
- We will prove later that Minimizing Cross Entropy Loss is equivalent to Maximizing the Likelihood of Training Data.
- Softmax at the outler layer typically needs cross entropy loss.
- “Distance” between two probalitiy distributions is the cross entropy loss.
- Softmax gives the observed probability distribution

Example

O/P: <+ve, neutral, –ve>

Obs: <0.8, 0.18, 0.02>

Tgt: <1, 0, 0>



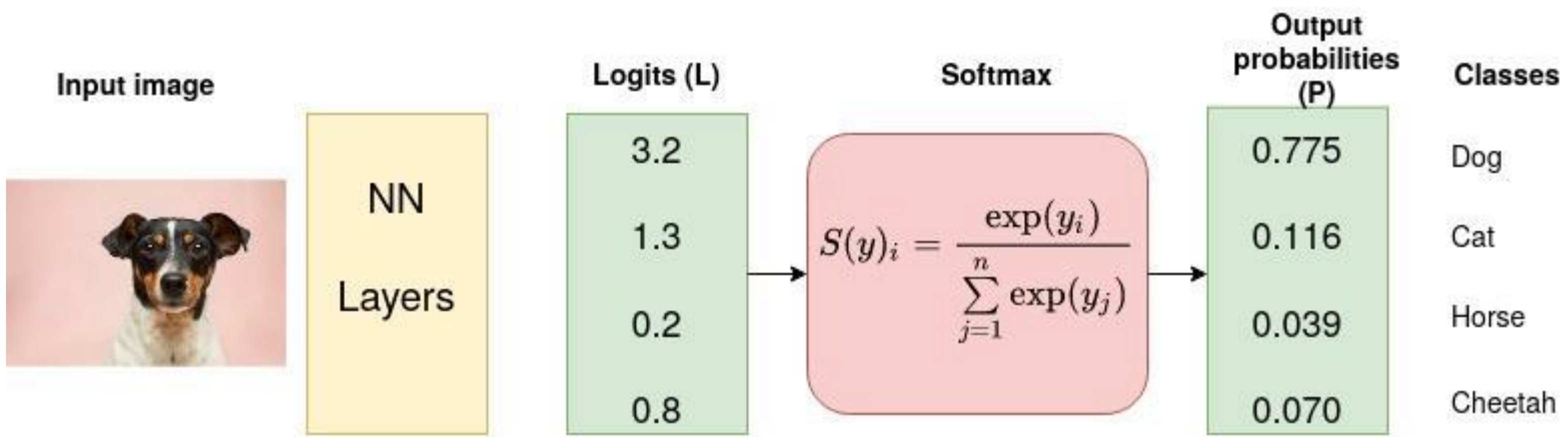
I loved the Movie

t_i:target, y_i: observed

$$CE = -\sum_{i=1}^{\#Classes} t_i \log y_i$$

$$= -[1.\log(0.8) + 0.\log(0.18) + 0.\log(0.12)]$$

Another Example: Image Recognition



Credit: https://medium.com/unpackai/cross-entropy-loss-in-ml-d9f22fc11fe0

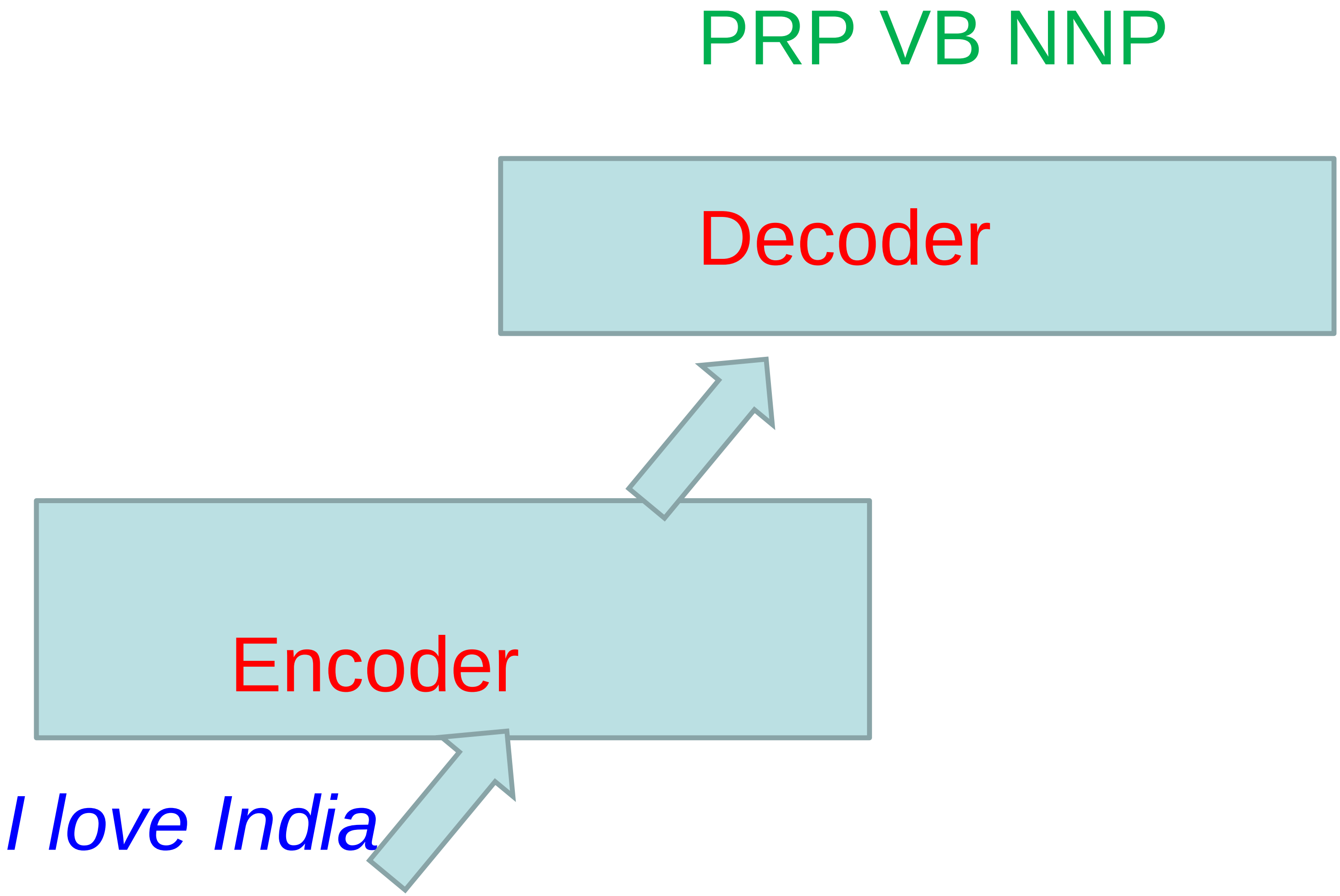
MLE: Maximize probability of training data

- *W*: Word sequence; *T*: Tag Sequence
- *P(W)*: probability of word sequence: **Language Model**
- *P(T|W):* probability of tag sequence given the word sequence

$$\arg\max_T [P(W,T)]$$

$$P(W,T) = P(W).P(T \mid W)$$

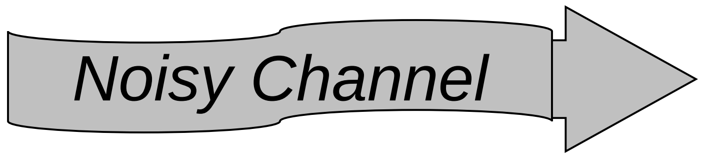
DNN is trained for MLE



Statistical POS tagging

Noisy Channel Model

W



T

$(w_n, w_{n-1}, \ldots, w_1)$

$(t_m, t_{m-1}, \ldots, t_1)$

Sequence *W* is transformed into sequence *T*

$$T^* = \underset{T}{\operatorname{argmax}}(P(T|W))$$

$$W^* = \underset{W}{\operatorname{argmax}}(P(T).P(W|T))$$

Argmax computation (1/2)

Best tag sequence
= T*
= argmax P(T|W)
= argmax P(T)P(W|T) (by Baye's Theorem)

$P(T) = P(t_0{=}\text{^}\ t_1 t_2 \ldots t_{n+1}{=}.)$

$= P(t_0)P(t_1|t_0)P(t_2|t_1t_0)P(t_3|t_2t_1t_0) \ldots P(t_n|t_{n-1}t_{n-2}\ldots t_0)P(t_{n+1}|t_nt_{n-1}\ldots t_0)$

$= P(t_0)P(t_1|t_0)P(t_2|t_1) \ldots P(t_n|t_{n-1})P(t_{n+1}|t_n)$

$$= \prod_{i=0}^{N+1} P(t_i|t_{i-1})$$ Bigram Assumption

Argmax computation (2/2)

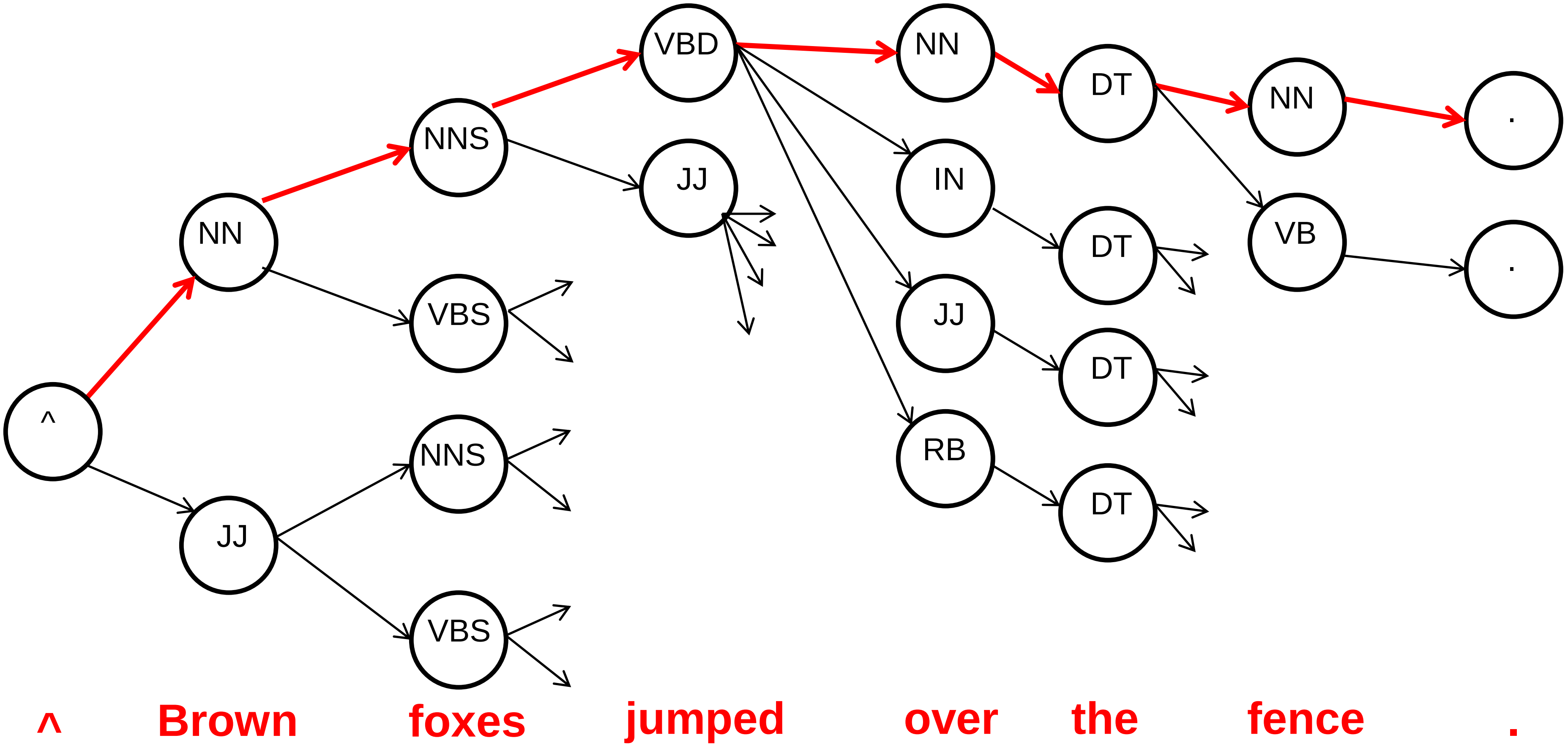
$$P(W|T) = P(w_0|t_0\text{-}t_{n+1})P(w_1|w_0t_0\text{-}t_{n+1})P(w_2|w_1w_0t_0\text{-}t_{n+1}) \ldots P(w_n|w_0\text{-}w_{n-1}t_0\text{-}t_{n+1})P(w_{n+1}|w_0\text{-}w_nt_0\text{-}t_{n+1})$$

Assumption: A word is determined completely by its tag. This is inspired by speech recognition

$$= P(w_o|t_o)P(w_1|t_1) \ldots P(w_{n+1}|t_{n+1})$$

$$= \prod_{i=0}^{n+1} P(w_i|t_i)$$

$$= \prod_{i=1}^{n+1} P(w_i|t_i) \quad \text{(Lexical Probability Assumption)}$$



Find the PATH with MAX Score.

What is the meaning of score?

CRF Based POS Tagging

Marathi

माणसाने उडण्याचा प्रयत्न केला

NN	**VG**	**NN**	**VBD**
B	**B**	**B**	**I**

Man tried flying

त्याने चालायला सुरुवात केली

PRP	**VINF**	**NN**	**VBD**
B	**B**	**B**	**I**

He started to walk

Harshada Gune, Mugdha Bapat, Mitesh Khapra and Pushpak Bhattacharyya, Verbs are where all the Action Lies: Experiences of Shallow Parsing of a Morphologically Rich Language, Computational Linguistics Conference (COLING 2010), Beijing, China, August 2010.

Decoding for the best Sequence

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} p_{\boldsymbol{\lambda}}(\boldsymbol{y}|\boldsymbol{x}) = \arg\max_{\boldsymbol{y}} \boldsymbol{\lambda} \cdot \boldsymbol{F}(\boldsymbol{y}, \boldsymbol{x})$$

$$p_{\boldsymbol{\lambda}}(\boldsymbol{Y}|\boldsymbol{X}) = \frac{\exp \boldsymbol{\lambda} \cdot \boldsymbol{F}(\boldsymbol{Y}, \boldsymbol{X})}{Z_{\boldsymbol{\lambda}}(\boldsymbol{X})} \qquad (1)$$

where

$$Z_{\boldsymbol{\lambda}}(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \exp \boldsymbol{\lambda} \cdot \boldsymbol{F}(\boldsymbol{y}, \boldsymbol{x})$$

$$\boldsymbol{F}(\boldsymbol{y}, \boldsymbol{x}) = \sum_{i} \boldsymbol{f}(\boldsymbol{y}, \boldsymbol{x}, i)$$

i ranges over the input positions

Representation

How to input text to neural net? Issue of REPRESENTATION

- Inputs have to be sets of numbers
  - We will soon see why

- These numbers form **REPRESENTATIONS**

- What is a good representation? At what granularity: words, n-grams, phrases, sentences

Issues

- What is a good representation? At what granularity: words, n-grams, phrases, sentences
- Sentence is important- (a) *I bank with SBI; (b) I took a stroll on the river bank; (c) this bank sanctions loans quickly*
- Each 'bank' should have a differengt representation
- We have to LEARN these representations

Principle behind representation

- Proverb: “A man is known by the company he keeps”
- Similalry: “A word is known/represented by the company it keeps”
- “Company” → Distributional Similarity

Representation: to learn or not learn?

- 1-hot representation does not capture many nuances, e.g., semantic similarity
  - But is a good starting point
- Collocations also do not fully capture all the facets
  - But is a good starting point

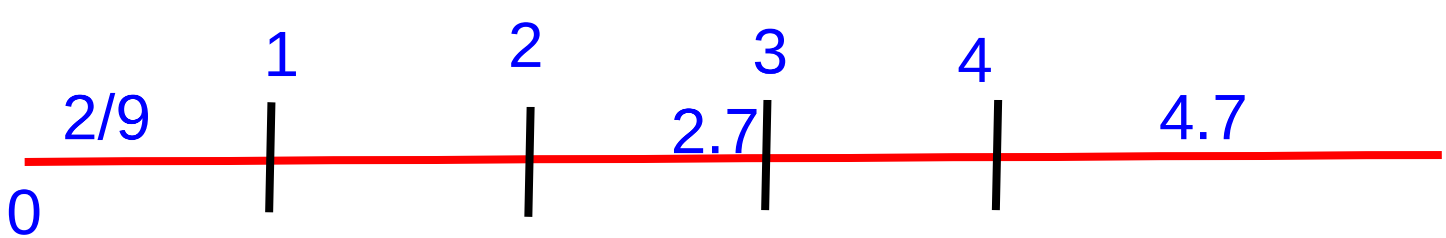
So learn the representation…

- Learning Objective

- ***MAXIMIZE CONTEXT PROBABILITY***

Foundations-1: Embedding

- Way of taking a discrete entity to a continuous space
- E.g., 1, 2, 3, 2.7, 2/9, $22^{1/2}$, … are numerical symbols
- But they are points on the real line
- Natural embedding
- Words' embedding not so intuitive!



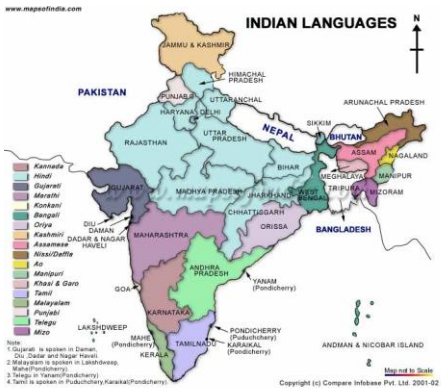
Foundations-2: Purpose of Embedding

- Enter geometric space
- Take advantage of “distance measures”- Euclidean distance, Riemannian distance and so on
- “Distance” gives a way of computing similarity

Foundations-3: Similarity and difference

- Recognizing similarity and difference- foundation of intelligence
- Lot of Pattern Recognition is devoted to this task (Duda, Hart, Stork, 2nd Edition, 2000)
- Lot of NLP is based on Text Similarity
- Words, phrases, sentences, paras and so on (verticals)
- Lexical, Syntactic, Semantic, Pragmatic (Horizontal)

Similarity study in MT



English:

This blanket is very soft

Hindi:

yaha kambal bahut naram hai

Bangla:

ei kambal ti khub naram <null>

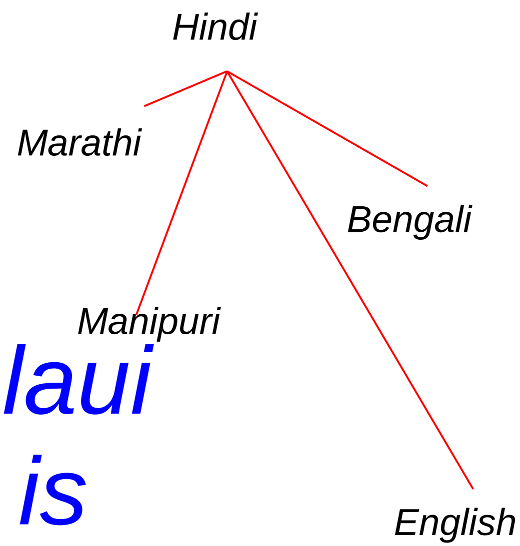
Marathi:

haa kambal khup naram aahe

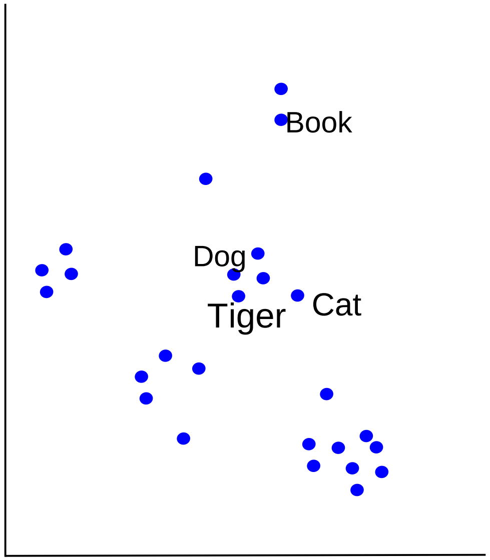
Manipuri:

kampor asi mon mon laui

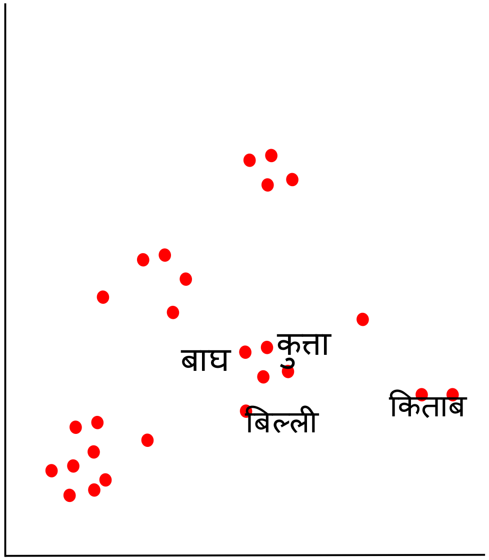
blanket this soft soft is



ISO-Metricity



English

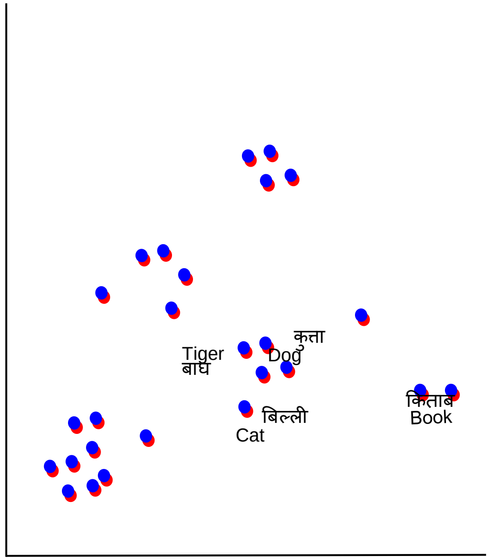


Hindi

Across Cross-lingual Mapping

This involves strong assumption that embedding spaces across languages are isomorphic, which is not true specifically for distance languages (Søgaard et al. 2018). However, without this assumption unsupervised NMT is not possible.

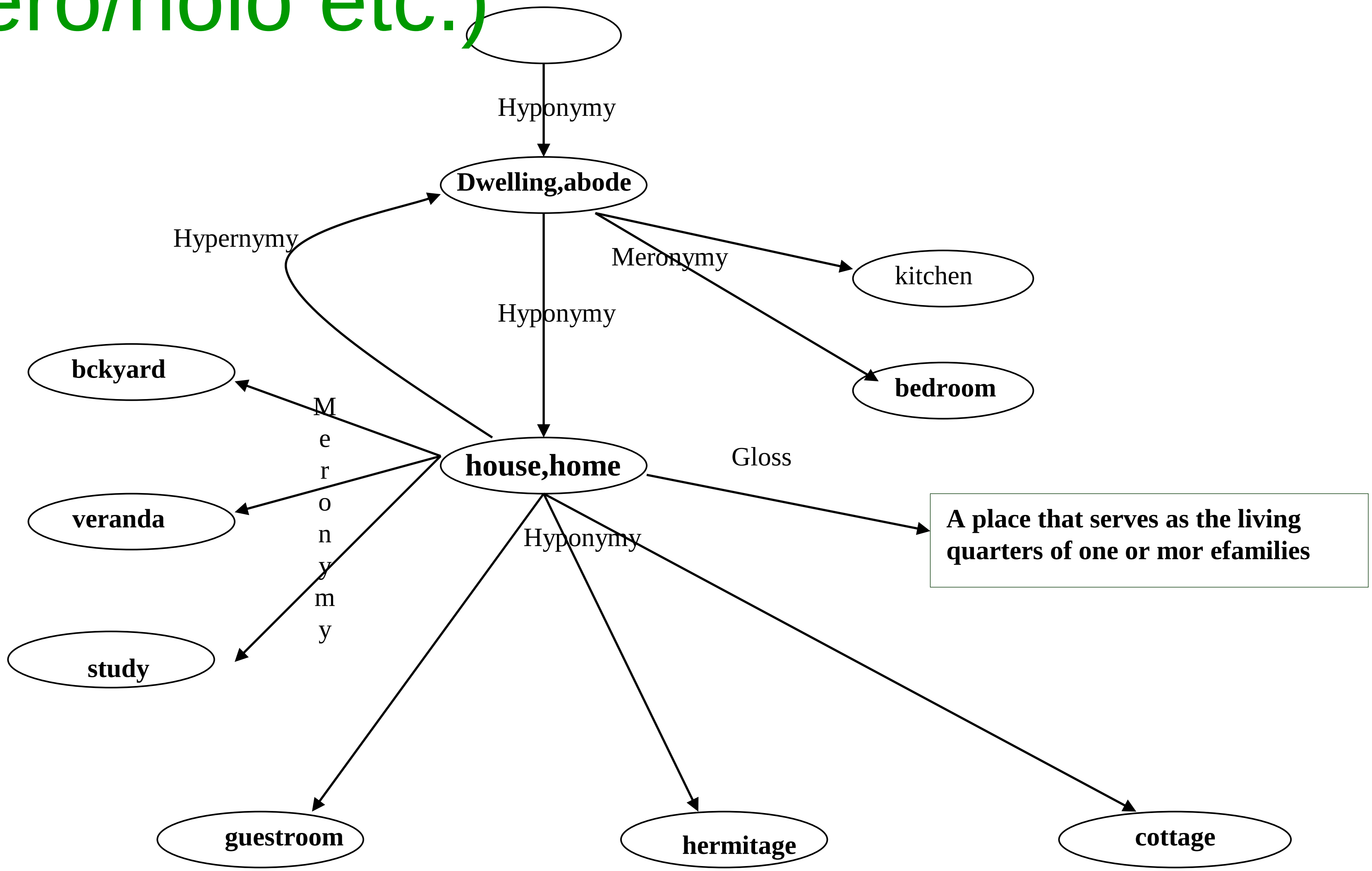
Søgaard, Anders, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. ACL



Foundations-4: Syntagmatic and Paradigmatic Relations

- Syntagmatic and paradigmatic relations
  - Lexico-semantic relations: synonymy, antonymy, hypernymy, mernymy, troponymy etc. **CAT is-a ANIMAL**
  - Coccurence: **CATS MEW**
- Wordnet: primarily paradigmatic relations
- ConceptNet: primarily Syntagmatic Relations

WordNet Sub-Graph with lexico-semantic relations (hyper/hypo, mero/holo etc.)



Lexical and Semantic relations in wordnet

1. Synonymy (e.g., *house, home*)
2. Hypernymy / Hyponymy (kind-of, e.g., *cat* ←→ *animal*)
3. Antonymy (e.g., *white and black*)
4. Meronymy / Holonymy (part of, e.g., *cat and tail*)
5. Gradation (e.g., *sleep*→*doze*→*wake up*)
6. Entailment (e.g., *snoring* → *sleeping*)
7. Troponymy (manner of, e.g., *whispering and talking*)

1, 3 and 5 are lexical (*word to word)*, rest are semantic (*synset to synset).*

‘Paradigmatic Relations’ and ‘Substitutability’

- Words in paradigmatic relations can substitute each other in the sentential context
- E.g., ‘The cat is drinking milk’ → ‘The animal is drinking milk’
- Substitutability is a foundational concept in linguistics and NLP

Foundations-5: Learning and Learning Objective

- Probability of getting the context words given the target should be maximized (skip gram)

- Probability of getting the target given context words should be maximized (CBOW)

Learning objective (skip gram)

$$J'(\theta) = \frac{1}{T} \prod_{t=1}^{T} \prod_{\substack{-m \le j \le m \\ j \ne 0}} p(w_{t+j} \mid w_t; \theta)$$

$$J(\theta) = -\frac{1}{T} \prod_{t=1}^{T} \prod_{\substack{-m \le j \le m \\ j \ne 0}} p(w_{t+j} \mid w_t; \theta)$$

$$Minimize \quad L = -\sum_{t=1}^{T} \sum_{\substack{-m \le j \le m \\ j \ne 0}} \log[p(w_{t+j} \mid w_t; \theta)]$$

Modelling *P(context word|input word)* *(1/2)*

- We want, say, *P('bark'|'dog')*
- Take the weight vector **FROM** 'dog' neuron **TO** projection layer (call this u_{dog})
- Take the weight vector **TO** 'bark' neuron **FROM** projection layer (call this v_{bark})
- When initialized u_{dog} and v_{bark} give the initial estimates of word vectors of 'dog' and 'bark'
- The weights and therefore the word vectors get fixed by back propagation

Modelling *P(context word|input word)* *(2/2)*

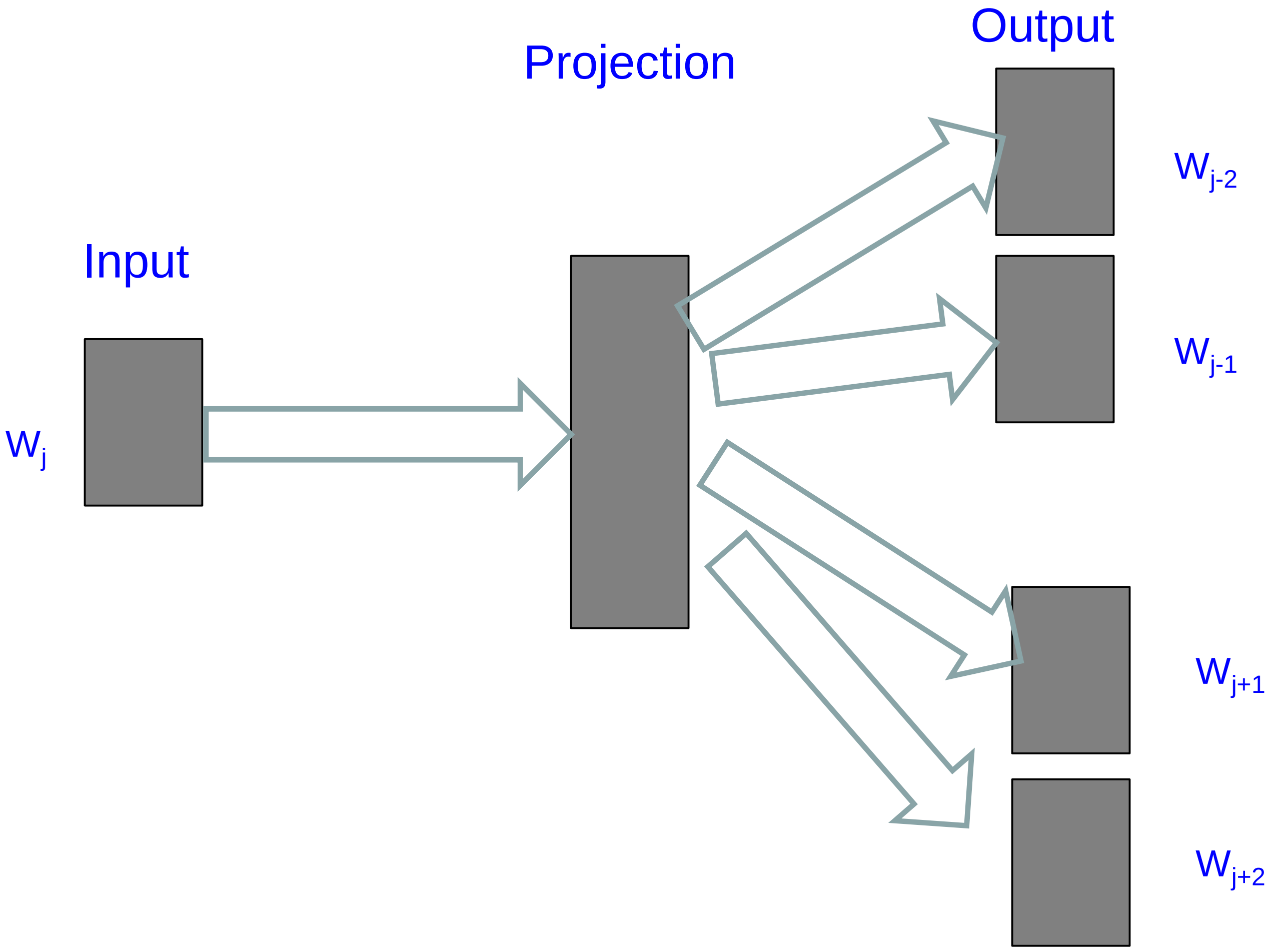
- To model the probability, first compute dot product of u_{dog} and v_{bark}
- Exponentiate the dot product
- Take softmax over all dot products over the whole vocabulary

$$P('bark'|'dog') = \frac{\exp(u_{dog}^T v_{bark})}{\sum_{v_k \varepsilon Vocabulary} \exp(u_{dog}^T v_k)}$$

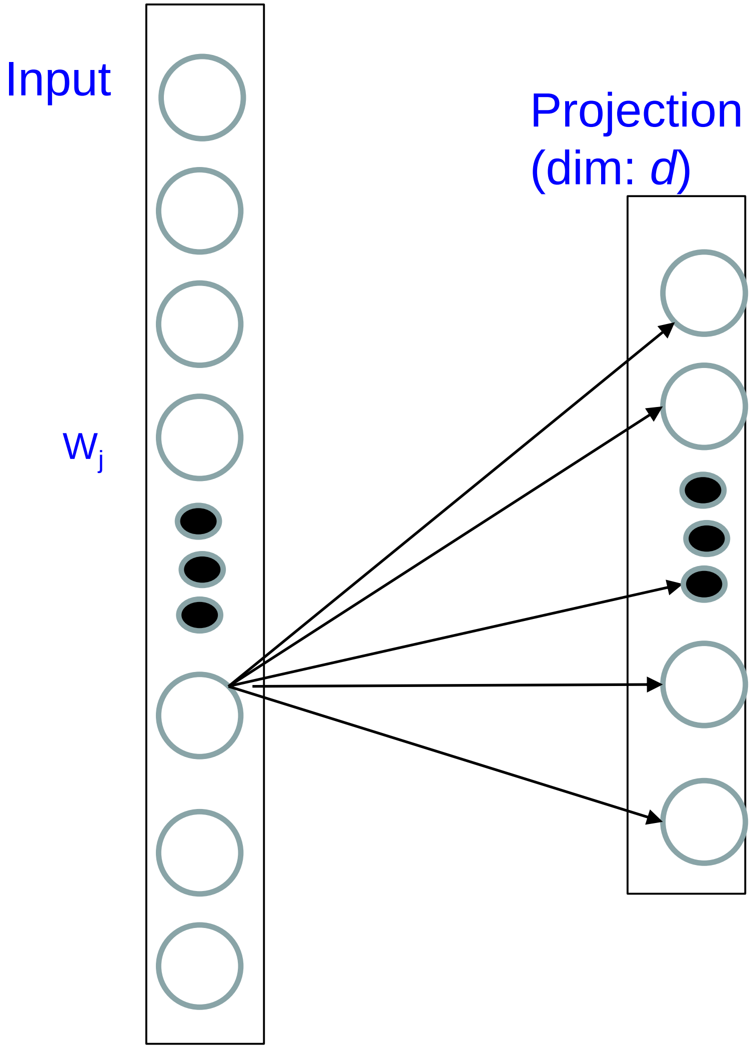
Exercise

- Why cannot you model *P(‘bark’|’dog’)* as the ratio of counts of <bark, dog> and <dog> in the corpus?
- Why this way of modelling probability through dot product of weight vectors of input and output words, exponentiation and soft-maxing works?

Modelling $p(w_{t+j}|w_t)$

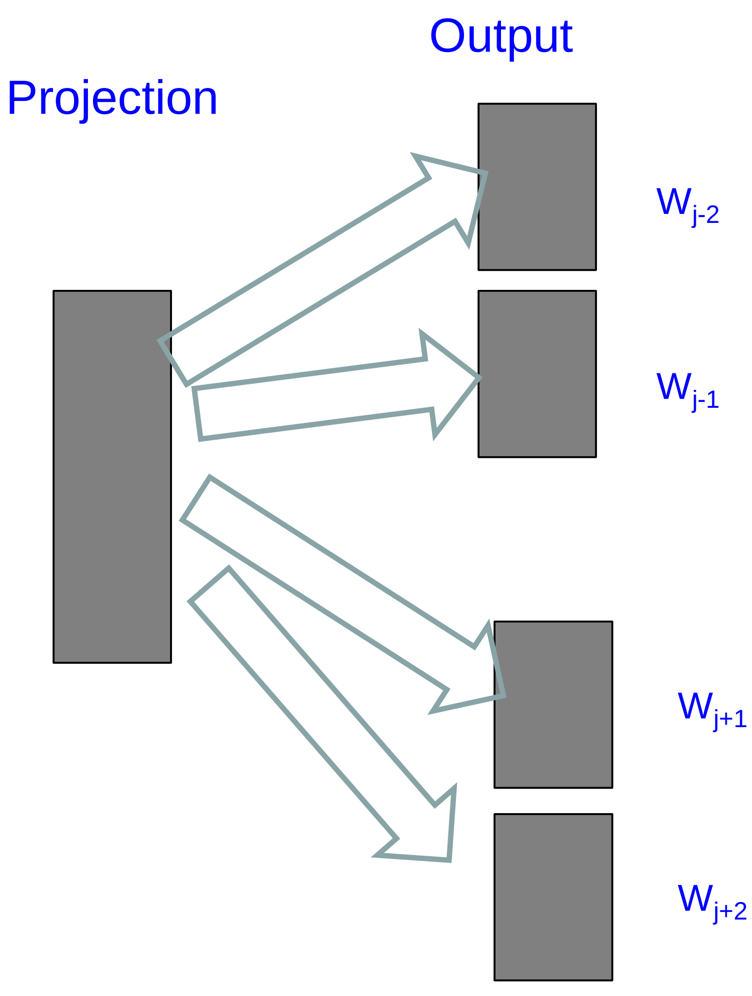


Input to Projection (shown for one neuron only)



- From each input neuron, a weight vector of dim *d*
- Input vector is of dim *V*, where V is the vocab size
- Input to projection we have a weight matrix *W* which is *V X d*
- Each row gives the weight vector of dim *d* REPRESENTING that word
- E.g., rows for ‘dog’, ‘cat, ‘lamp’, ‘table’ etc.

Projection to output



- From the whole projection layer a weight vector of dim *d* to each neuron in each compartment, where the compartment represents a context word
- Each fat arrow is a *d X V* matrix

Linguistic foundation of word representation

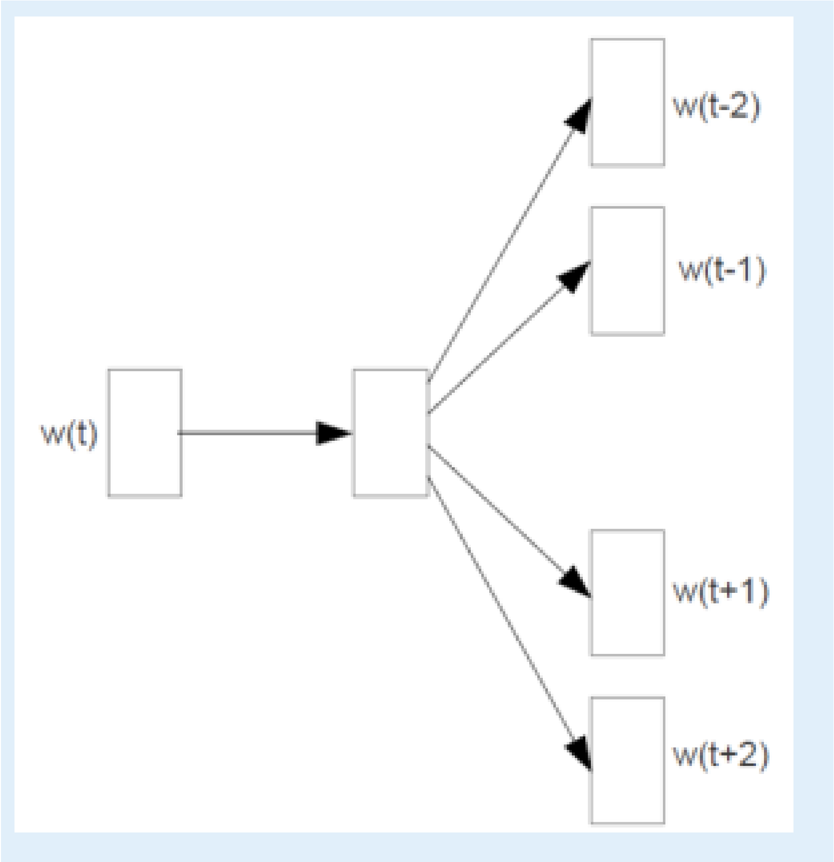
“Linguistics is the eye”: Harris Distributional Hypothesis

- Words with similar distributional properties have similar meanings. (Harris 1970)

- 1950s: Firth- “A word is known by the company its keeps”

- Model **differences** in meaning rather than the proper meaning itself

"Computation is the body": Skip gram- predict context from word



For CBOW:

Just reverse the Input-Ouput

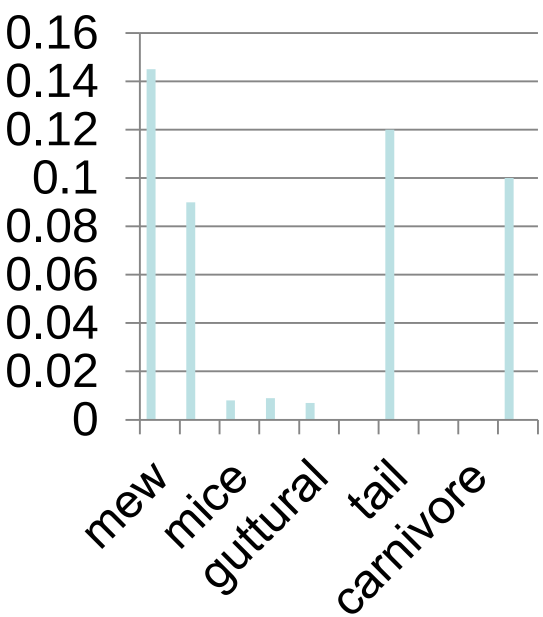
Dog – Cat - Lamp

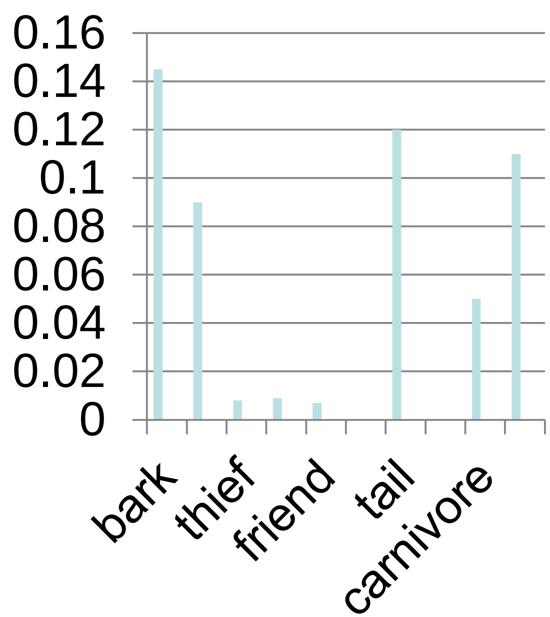
{bark, police, thief, vigilance, faithful, friend, animal, milk, carnivore)

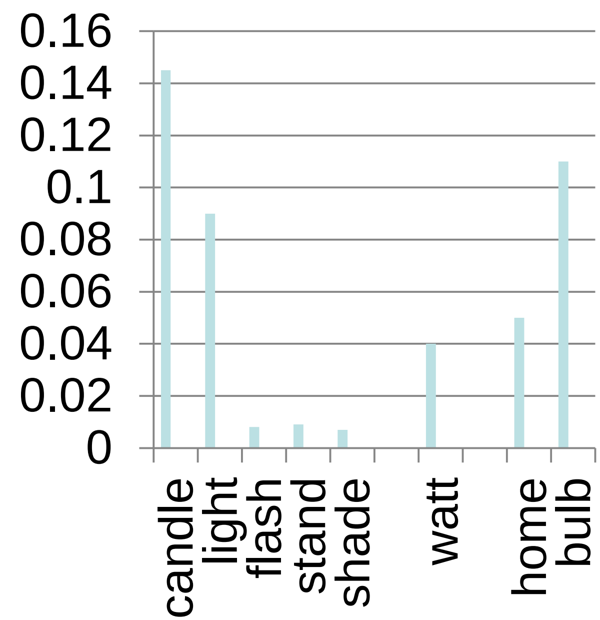
{mew, comfort, mice, furry, guttural, purr, carnivore, milk}

{candle, light, flash, stand, shade, Halogen}

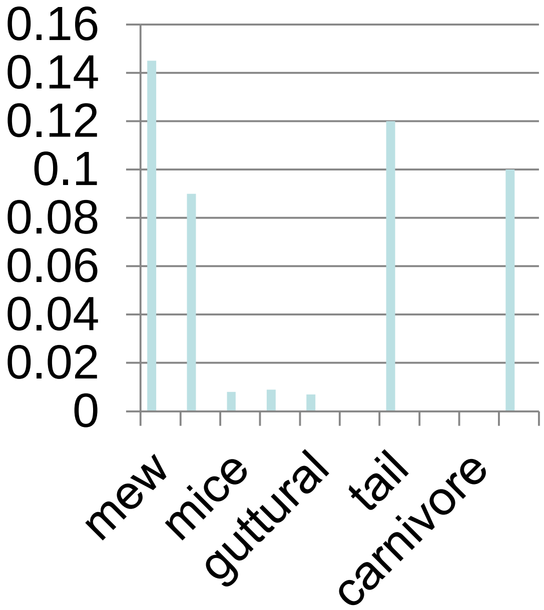
Probability distributions of context words



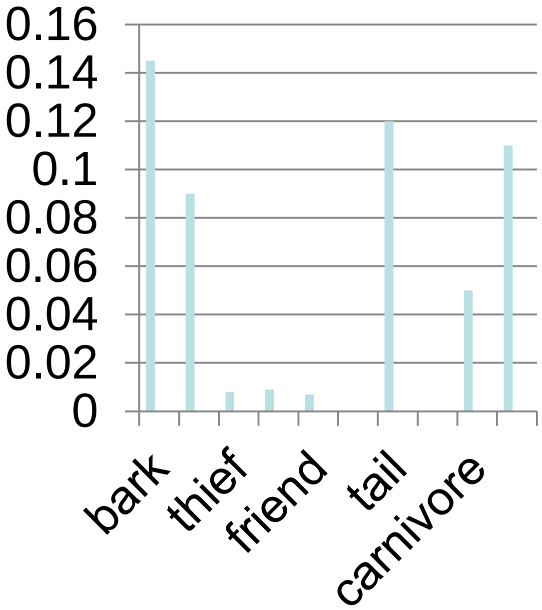




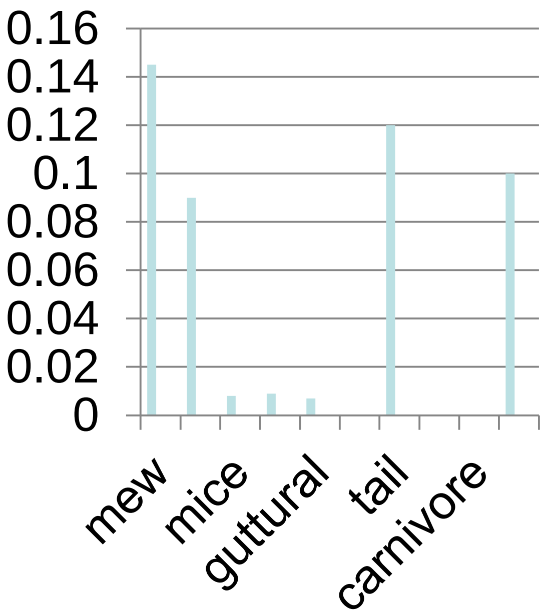
Probability distributions of context words



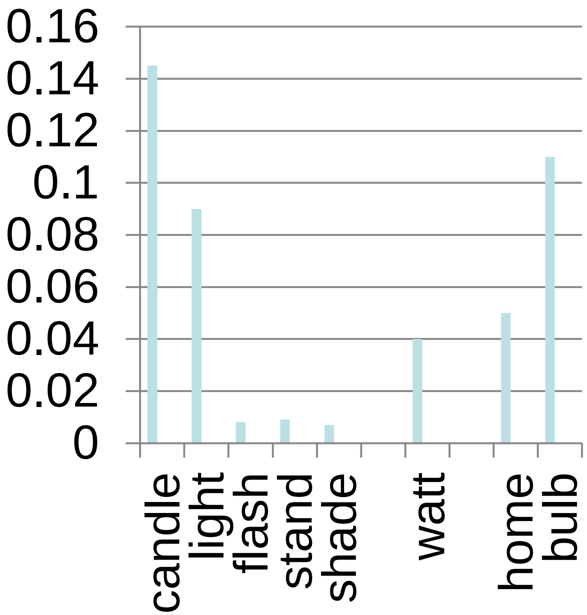
$\cong$



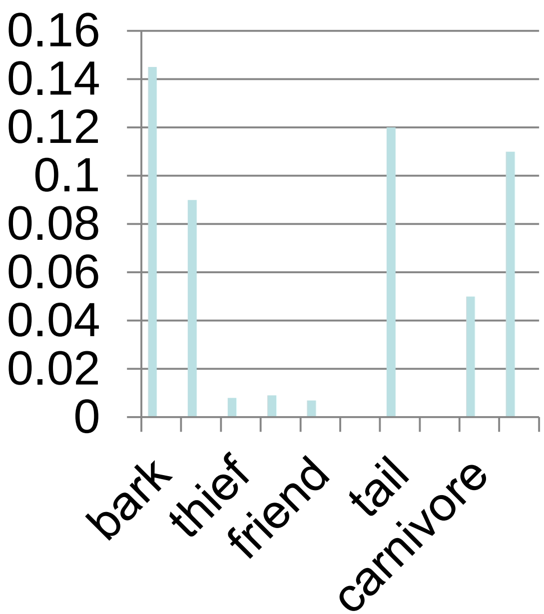
Probability distributions of context words



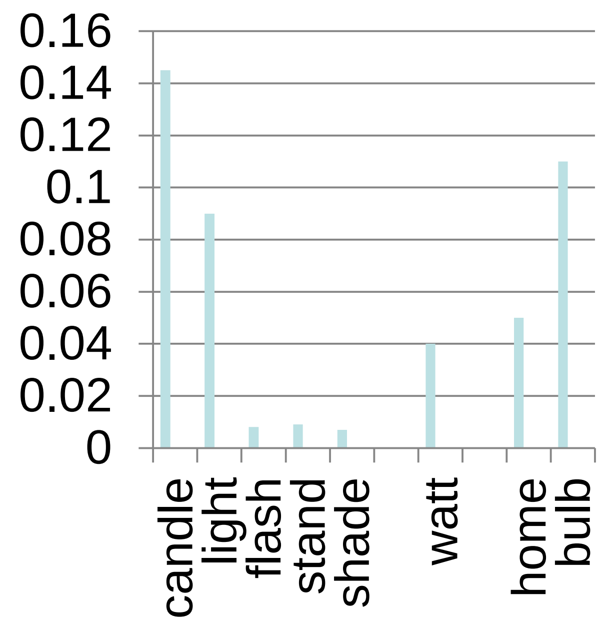
≠



Probability distributions of context words



$\neq$



Test of representation

- **Similarity**
  - ‘Dog’ more similar to ‘Cat’ than ‘Lamp’, because
  - Input- vector(‘dog’), output- vectors of associated words
  - More similar to output from vector(‘cat’) than from vector(‘lamp’)

“Linguistics is the eye, Computation is the body”

The encode-decoder deep learning network is nothing but

the ***implementation*** of

Harris’s Distributional Hypothesis

Distributed Representations of words

- Also known as word vectors, word embeddings, etc.
- Primarily, they are vectors in n-dimensional space
- Try to model meaning of word

Harris Distributional Hypothesis

- Words with similar distributional properties have similar meanings. (Harris 1970)
- Harris does mentions that distributional approaches can model differences in meaning rather than the proper meaning itself