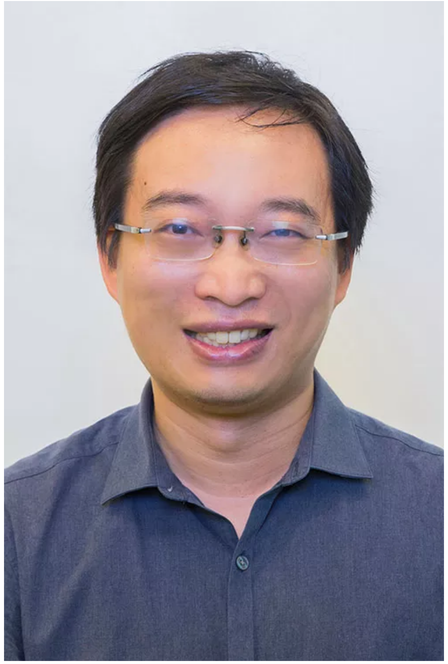
50.040
Natural Language Processing

Lu Wei

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Instructor



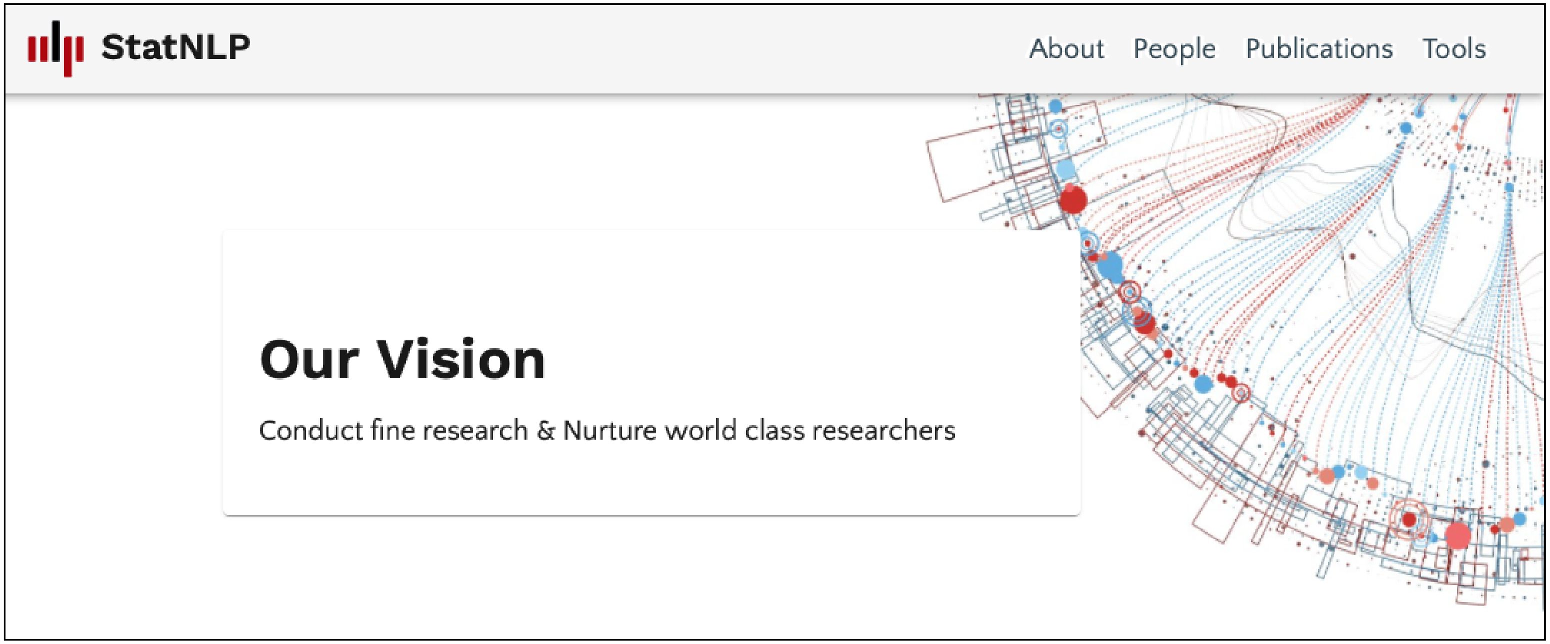
Lu Wei

1.302.10

luwei@sutd.edu.sg

www.statnlp.org

NLP Research



AI SINGAPORE

Alibaba Group 阿里巴巴集团

BOEING

DSO NATIONAL LABORATORIES

SUTD-MIT INTERNATIONAL DESIGN CENTRE (IDC) Innovation Ready Design

国家自然科学基金委员会 1986 NATIONAL NATURAL SCIENCE FOUNDATION OF CHINA

Ministry of Education SINGAPORE

NATIONAL RESEARCH FOUNDATION PRIME MINISTER'S OFFICE SINGAPORE

MINDEF SINGAPORE

字节跳动 ByteDance

NLP Research


Cheng Liying — PhD Student
Research Interests: Text Generation


Guo ZhiJiang — PhD Student
Research Interests: Syntatic Parsing; Semantic Parsing


Jie Zhanming — PhD Student
Research Interests: Semantic Parsing; Named Entity Recognition


Joel Ong — PhD Student
Research Interests: Natural Language Parsing; Machine Learning in Finance; Reinforcement Learning


Li Hao — PhD Student
Research Interests: Machine Learning; Artificial Intelligence; Natural Language Processing


Li Haoran — PhD Student
Research Interests: Machine Learning; Graph Neural Networks; Natural Language Processing


Perry Lam — PhD Student
Research Interests: Speech Generation; Conversational Information Extraction; Computational Psycholinguistics


Sun Xiaobing — PhD Student
Research Interests: Target Sentiment Analysis; Deep Learning Interpretability


Thilini — PhD Student
Research Interests: Visual Language Tasks; Visual Question & Answering


Xu Lu — PhD Student
Research Interests: Graph Neural Network; Low Resource Task; Relation & Information Extraction


Zhang Yan — PhD Student
Research Interests: Graph Neural Networks; Knowledge Representation & Reasoning


Zou Yanyan — PhD Student
Research Interests: Natural Language Understanding; Semantic Parsing; Text Summarization


Nan Guoshun — Postdoc
Research Interests: Named Entity Recognition; Relation Extraction; Question & Answering


Qiao Rui — Research Asst.
Research Interests: Machine Learning; Structured Prediction


Zhuang Boyuan — Research Asst.
Research Interests: Discourse Analysis; Implicit Discourse Analysis; Neural Network Interpretability


Liao Jing — Visiting Student
Research Interests: Sentiment Analysis; Intelligent Design; Human Centered Computing


Ivan Sekulic — Research Intern
Research Interests: Machine Learning; Natural Language Processing


Wang Huiming — Research Intern
Research Interests: Named Entity Recognition; Semantic Parsing; Machine Learning


Zhou Yuxiang — Visiting Student
Research Interests: Text Classification; Sentiment Analysis; Information Extraction


Cheng Juan — Postdoc
Research Interests: Sentiment Analysis; Relation Extraction; Deep Learning Interpretability


Shi Ke — Research Asst.
Research Interests: Named Entity Recognition; Relation Extraction; Question & Answering


Lu Wei — Associate Professor
Research Interests: Machine Learning; Artificial Intelligence; Natural Language Processing

Major NLP Conferences

Course Plan

50.040 Natural Language Processing
(Undergraduate)
Summer 2020

Instructor	Lu Wei	luwei@sutd.edu.sg	1.302.10
Teaching Assistant	Haoran Li	haoran2_li@mymail.sutd.edu.sg	1.417
Office Hours	Lu Wei Haoran Li	Thursday 15:30 - 16:30pm, or by appointment Friday 14:00 - 15:00pm, or by appointment	1.302.10 1.417
Lessons	Every Week	Mon 10:30 – 12:30 Thu 13:30 – 15:30	
Classroom	Think Tank 24 (2.503)		
Prerequisites	1. 01.112 Machine Learning or 01.113 Statistical and Machine Learning and 2. A good foundation in: 1) programming, 2) design and analysis of algorithms, 3) mathematics including linear algebra, calculus, optimization, probability, and statistics.		
Assessment	Homework (coding) (30%), Homework (written) (20%), Mini Project (20%), Final Project (30%)		

Prerequisites

Machine Learning

Algorithms

Linear Algebra

Deep Learning

Prob & Stats

Optimization

Programming

Assessment

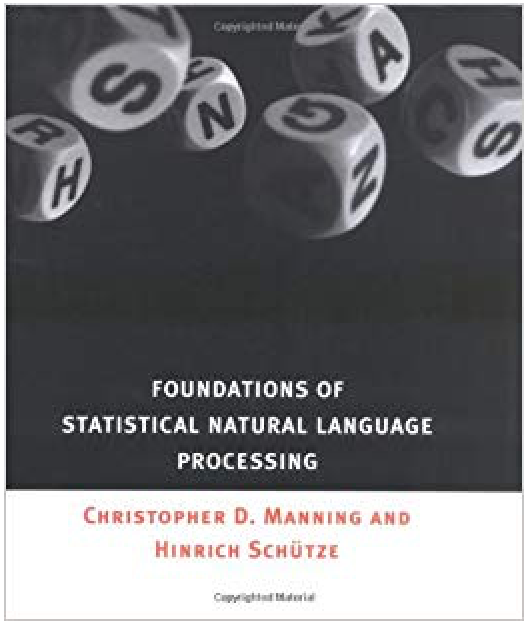
Item	Weightage	Remark
Homework	30%	Biweekly
Mini Project	20%	Individual
Written Homework	20%	6-weekly
Final Project	30%	Group (2-3)

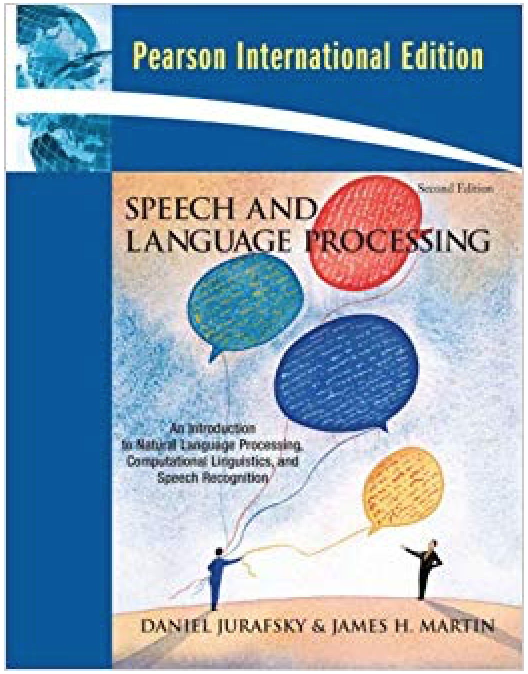
Learning Objectives

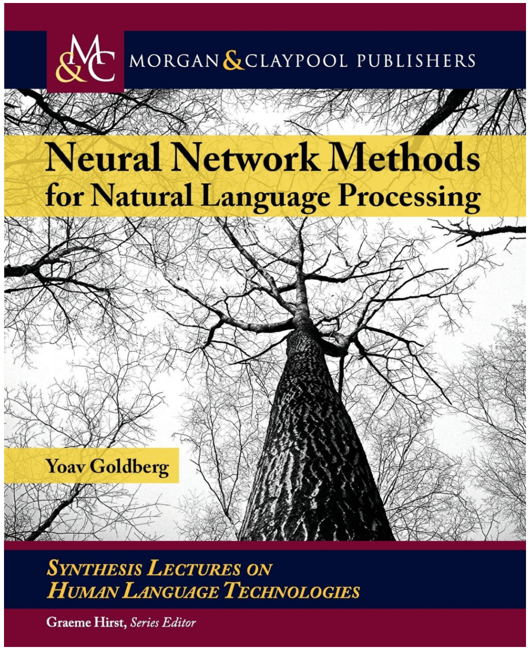
By the end of the course, students will be able to:

1. Explain the fundamental tasks within NLP
2. Explain possible algorithms as solutions to NLP tasks
3. Implement the algorithms used for various NLP tasks
4. Design novel algorithms for solving new NLP tasks, and use existing NLP technologies for solving real problems

Textbooks


FOUNDATIONS OF
STATISTICAL NATURAL LANGUAGE
PROCESSING
CHRISTOPHER D. MANNING AND
HINRICH SCHÜTZE


Pearson International Edition
Second Edition
SPEECH AND
LANGUAGE PROCESSING
An Introduction
to Natural Language Processing,
Computational Linguistics, and
Speech Recognition
DANIEL JURAFSKY & JAMES H. MARTIN


MORGAN & CLAYPOOL PUBLISHERS
Neural Network Methods
for Natural Language Processing
Yoav Goldberg
SYNTHESIS LECTURES ON
HUMAN LANGUAGE TECHNOLOGIES
Graeme Hirst, Series Editor

Honour Code

You are encouraged to form study groups to work on the homework together. However, please write out the solutions on your own, without referring to notes from other students, online solutions or answers from past courses. Reproducing the solution from scratch will help you to find out if you fully understand the material. Please also list your group members on the first page of your submission.

What is Natural Language Processing

Tasks in NLP

POS Tagging
Chunking
Document Classification
Information Extraction
Syntactic Parsing
Semantic Parsing
Natural Language Generation
Machine Translation
Sentiment Analysis
Coreference Resolution
Question Answering

Tasks in NLP

POS Tagging
Chunking
Document Classification
Information Extraction
Syntactic Parsing
Semantic Parsing
Natural Language Generation
Machine Translation
Sentiment Analysis
Coreference Resolution
Question Answering

Supervised

Word Clusters
GloVe, word2vec
Topic Modeling
Language Modeling
ELMo, BERT

Unsupervised

Part-of-Speech Tagging

A	**N**	**V**	**D**	**N**
Fruit	*flies*	*like*	*a*	*banana*

StatNLP.org

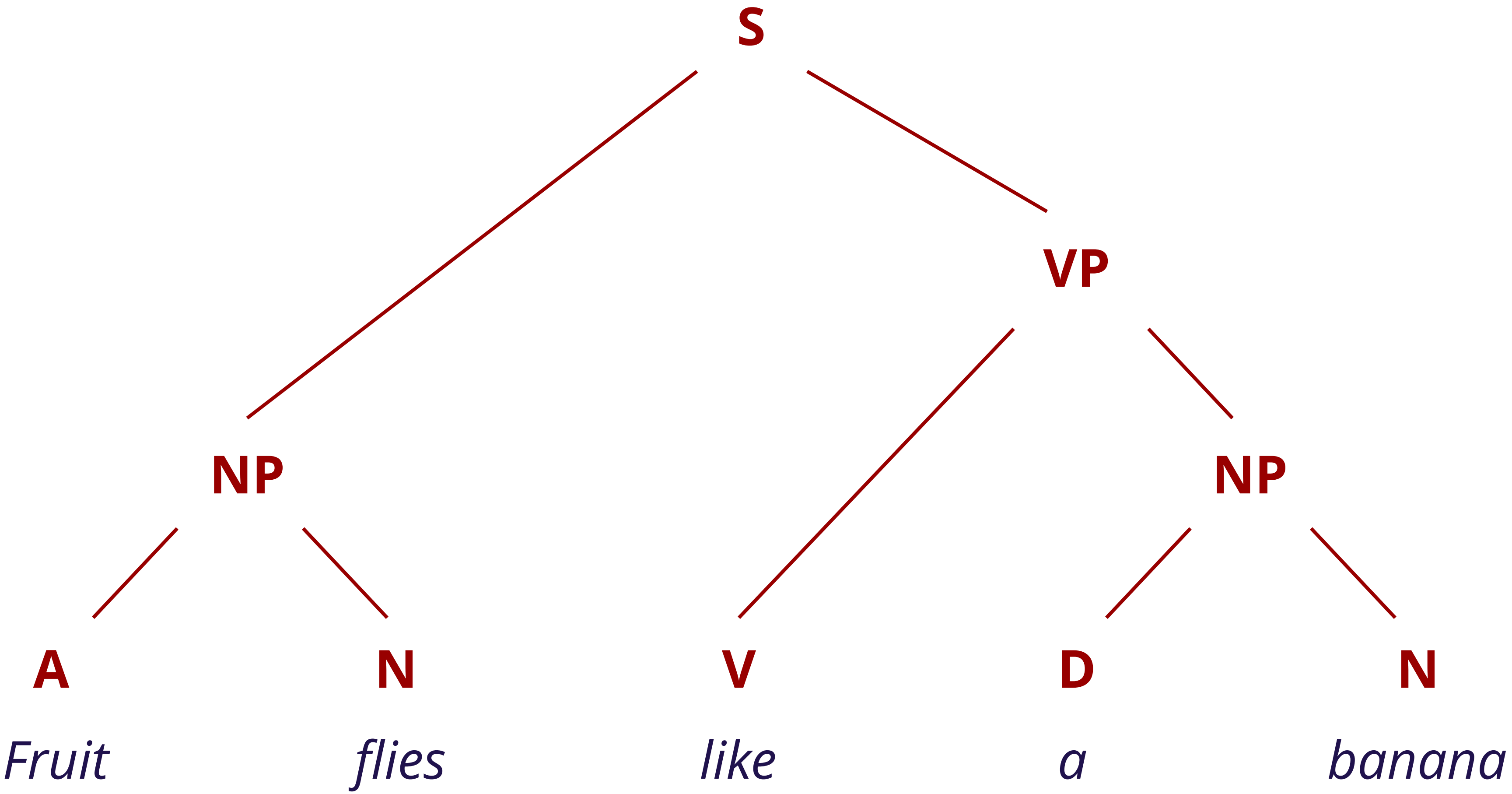
Noun-Phrase Chunking

NP NP

Fruit *flies* *like* *a* *banana*

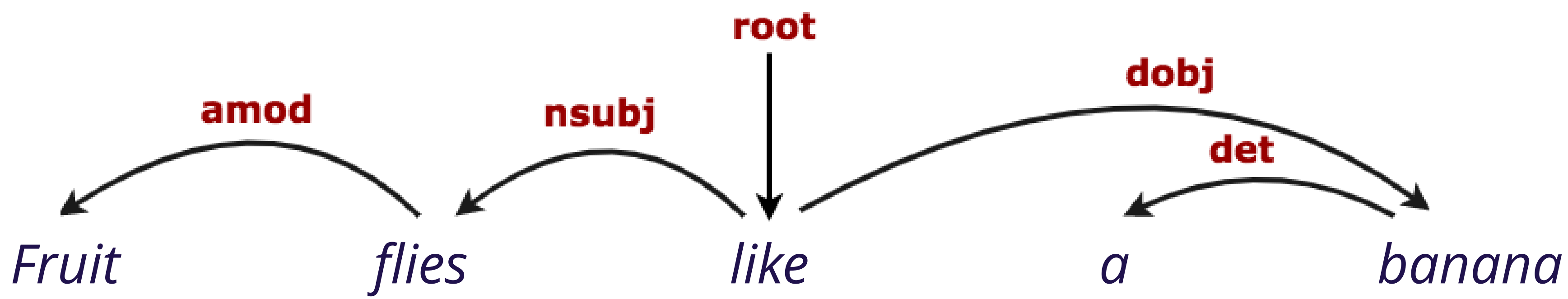
StatNLP.org

Constituency Parsing



StatNLP.org

Dependency Parsing



StatNLP.org

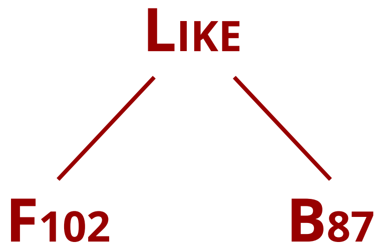
Semantic Parsing

LIKE(F102, B87)

Fruit *flies* *like* *a* *banana*

StatNLP.org

Semantic Parsing



Fruit *flies* *like* *a* *banana*

StatNLP.org

Sentiment Analysis

(neutral) **(positive)**

Fruit *flies* *like* *a* *banana*

StatNLP.org

Tasks in NLP

POS Tagging
Chunking
Document Classification
Information Extraction
Syntactic Parsing
Semantic Parsing
Natural Language Generation
Machine Translation
Sentiment Analysis
Coreference Resolution
Question Answering

Supervised

Many (Supervised) Problems in NLP are Structured Prediction Problems!

Tasks in NLP

POS Tagging

Chunking

Document Classification

Information Extraction

Syntactic Parsing

Semantic Parsing

Natural Language Generation

Machine Translation

Sentiment Analysis

Coreference Resolution

Question Answering

Supervised

Question

How to Build a Document Classification Model?

Word Representation

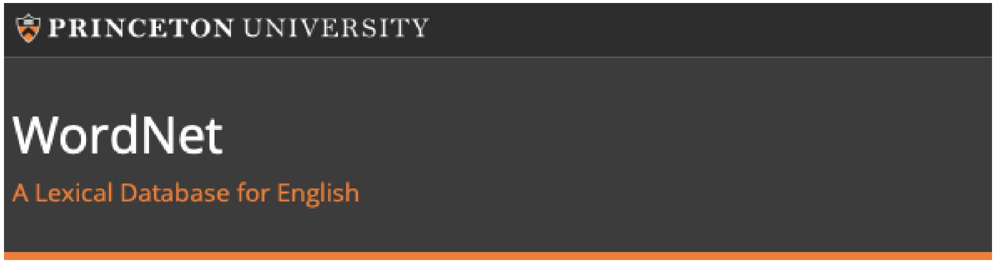
$$\begin{array}{ccc} \texttt{cat} & \texttt{eat} & \texttt{dog} \\ \begin{pmatrix} 0 \\ \color{red}{1} \\ 0 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 0 \end{pmatrix} & \begin{pmatrix} 0 \\ 0 \\ \color{red}{1} \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 0 \end{pmatrix} & \begin{pmatrix} 0 \\ 0 \\ 0 \\ \color{red}{1} \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 0 \end{pmatrix} \end{array}$$

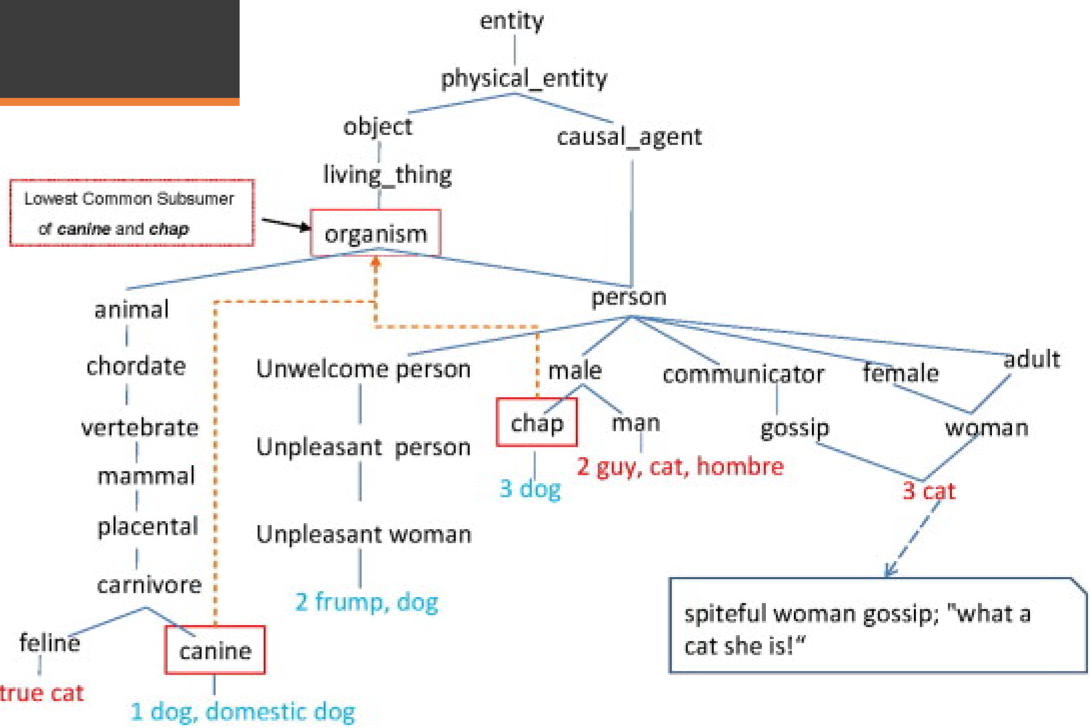
1-hot representations are sparse, which cannot capture *semantic* relations between words!

Need a way to capture *semantics*!

WordNet

https://wordnet.princeton.edu





Wei, Tingting, et al. "A semantic approach for text clustering using WordNet and lexical chains." *Expert Systems with Applications* 42.4 (2015): 2264-2275.

WordNet

Limitations:

Only for English

Incomplete

Subjective

Expensive to Maintain

Need a way to capture *semantics* ***automatically***!

Question

How to Automatically Group Similar Data Points?

Tasks in NLP

Word Clusters

GloVe, word2vec

Topic Modeling

Language Modeling

ELMo, BERT

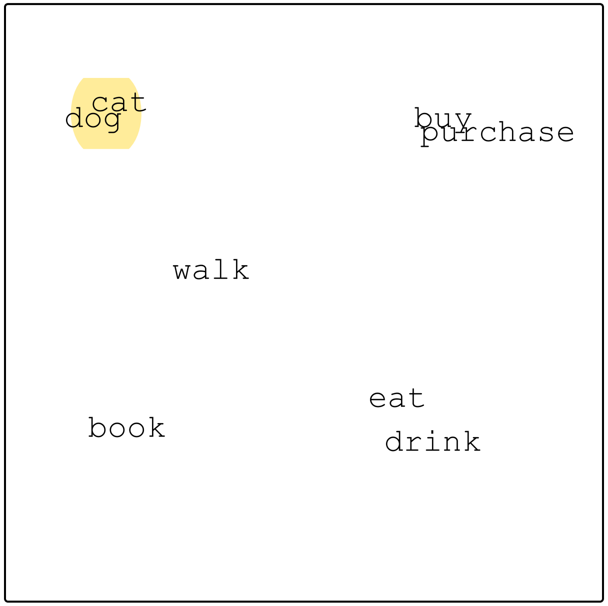
Unsupervised

Word Embeddings

$$\begin{array}{ccc} \texttt{cat} & \texttt{eat} & \texttt{dog} \\ \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ . \\ . \\ . \\ 0 \\ 0 \end{pmatrix} & \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ . \\ . \\ . \\ 0 \\ 0 \end{pmatrix} & \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ . \\ . \\ . \\ 0 \\ 0 \end{pmatrix} \end{array}$$

Word Embeddings

$$
\begin{array}{ccc}
\texttt{cat} & \texttt{eat} & \texttt{dog} \\
\begin{pmatrix} 0.232 \\ -0.903 \\ 0.213 \\ 0.239 \\ -0.679 \\ 0.923 \end{pmatrix} &
\begin{pmatrix} -0.932 \\ -0.903 \\ 0.213 \\ 0.590 \\ 0.609 \\ 0.388 \end{pmatrix} &
\begin{pmatrix} 0.190 \\ -0.887 \\ 0.193 \\ 0.257 \\ -0.708 \\ 0.906 \end{pmatrix}
\end{array}
$$



Tasks in NLP

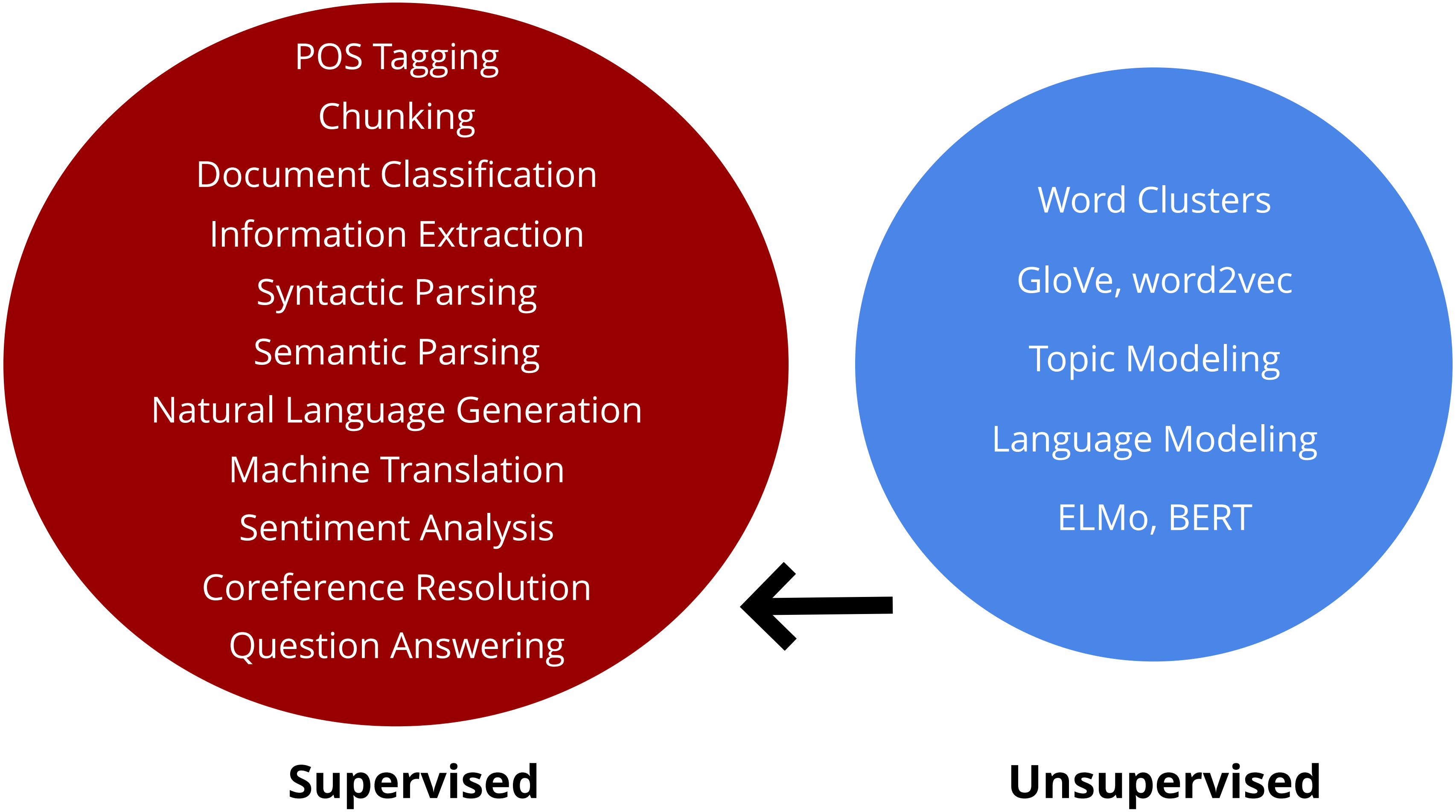
POS Tagging
Chunking
Document Classification
Information Extraction
Syntactic Parsing
Semantic Parsing
Natural Language Generation
Machine Translation
Sentiment Analysis
Coreference Resolution
Question Answering

Supervised

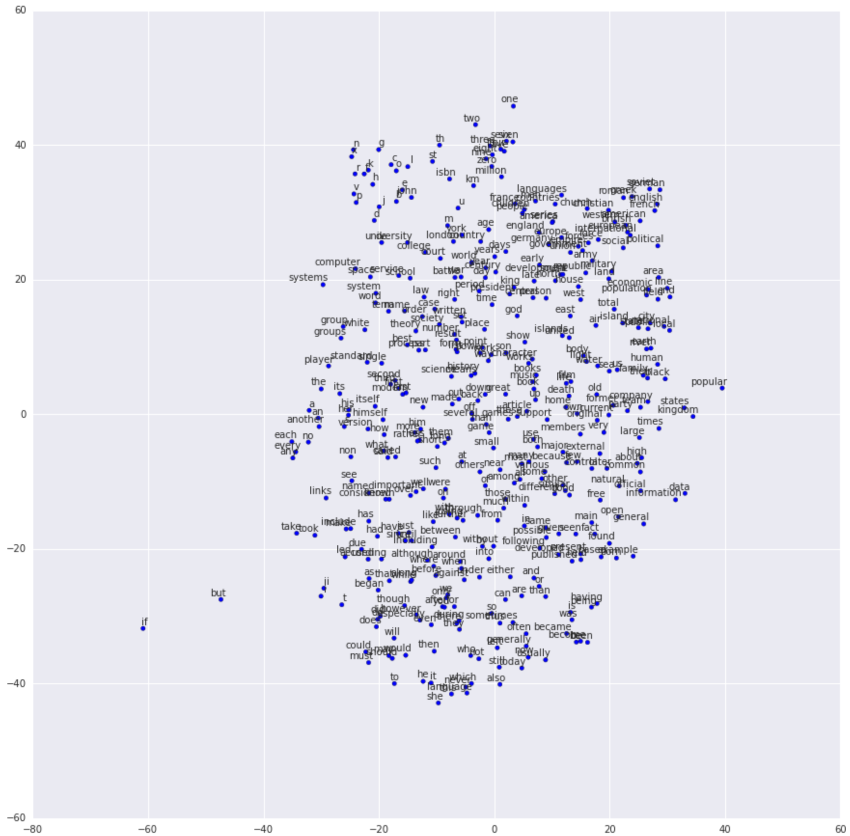
Word Clusters
GloVe, word2vec
Topic Modeling
Language Modeling
ELMo, BERT

Unsupervised

Tasks in NLP


POS Tagging
Chunking
Document Classification
Information Extraction
Syntactic Parsing
Semantic Parsing
Natural Language Generation
Machine Translation
Sentiment Analysis
Coreference Resolution
Question Answering
Word Clusters
GloVe, word2vec
Topic Modeling
Language Modeling
ELMo, BERT
Supervised
Unsupervised

Word Embeddings



Maps words in a vocabulary to vectors of continuous real numbers, and these vectors are in a lower-dimensional space

Question

How to Learn Word Embeddings?

Distributional Semantics

You shall know a word by the company it keeps

- Firth, J. R. 1957

Distributional Semantics

You shall know a word by the company it keeps

- Firth, J. R. 1957

Two Major Approaches to Learning Word Embeddings

Count-based

Prediction-based

Co-Occurence Matrix

$$
\begin{array}{cc}
 & \begin{array}{cccccccc} & \texttt{cat} & & & \texttt{eat} & & \texttt{dog} & \end{array} \\
\begin{array}{c} \\ \texttt{cat} \\ \\ \\ \texttt{eat} \\ \\ \texttt{dog} \\ \\ \end{array} &
\begin{bmatrix}
0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\
2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0
\end{bmatrix}
\end{array}
$$

Co-Occurence Matrix

$$
\begin{array}{cc}
 & \begin{array}{cccccccc} & \texttt{cat} & & & \texttt{eat} & & \texttt{dog} & \end{array} \\
\begin{array}{c} \\ \texttt{cat} \\ \\ \\ \texttt{eat} \\ \\ \texttt{dog} \\ \\ \end{array} &
\begin{bmatrix}
0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\
2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0
\end{bmatrix}
\end{array}
$$

Co-Occurence Matrix

$$
\begin{array}{c c}
 & \begin{array}{cccccccc} & \texttt{cat} & & & \texttt{eat} & & \texttt{dog} & \end{array} \\
\begin{array}{c} \\ \texttt{cat} \\ \\ \\ \texttt{eat} \\ \\ \texttt{dog} \\ \\ \end{array} &
\begin{bmatrix}
0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\
2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0
\end{bmatrix}
\end{array}
$$

Still very high dimensional! Still very sparse!

Dimensionality reduction: for example - SVD!

Co-Occurence Matrix

$$
\begin{array}{c|c}
 & \begin{array}{cccccccc} & \texttt{cat} & & & \texttt{eat} & & \texttt{dog} & \end{array} \\
\begin{array}{c} \\ \texttt{cat} \\ \\ \\ \texttt{eat} \\ \\ \texttt{dog} \\ \\ \end{array} &
\begin{bmatrix}
0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\
2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0
\end{bmatrix}
\end{array}
$$

$$
\begin{pmatrix} 0.232 \\ -0.903 \\ 0.213 \\ 0.239 \\ -0.679 \\ 0.923 \end{pmatrix}
\qquad
\begin{pmatrix} -0.932 \\ -0.903 \\ 0.213 \\ 0.590 \\ 0.609 \\ 0.388 \end{pmatrix}
\qquad
\begin{pmatrix} 0.190 \\ -0.887 \\ 0.193 \\ 0.257 \\ -0.708 \\ 0.906 \end{pmatrix}
$$

Matrix decomposition / dimension reduction methods will project them into a lower-dimensional space

Alternative Matrices

$$\begin{array}{cc}
 & \begin{array}{cccccccc} & \text{cat} & & & \text{eat} & & \text{dog} & \end{array} \\
\begin{array}{c} \\ \text{cat} \\ \\ \\ \text{eat} \\ \\ \text{dog} \\ \\ \end{array} &
\begin{bmatrix}
0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\
2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0
\end{bmatrix}
\end{array}$$

Binary Coccurence Matrix, TF-IDF Matrix, PMI Matrix, Shifted PMI Matrix ...

Distributional Semantics

You shall know a word by the company it keeps

- Firth, J. R. 1957

Two Major Approaches to Learning Word Embeddings

Count-based ✔

Prediction-based

Distributional Semantics

Count based	Prediction based

- Compute the statistics of how often each word co-occurs/interacts with its neighboring words in a large text corpus.
- Learn to map these count-statistics down to a small, dense vector for each word.

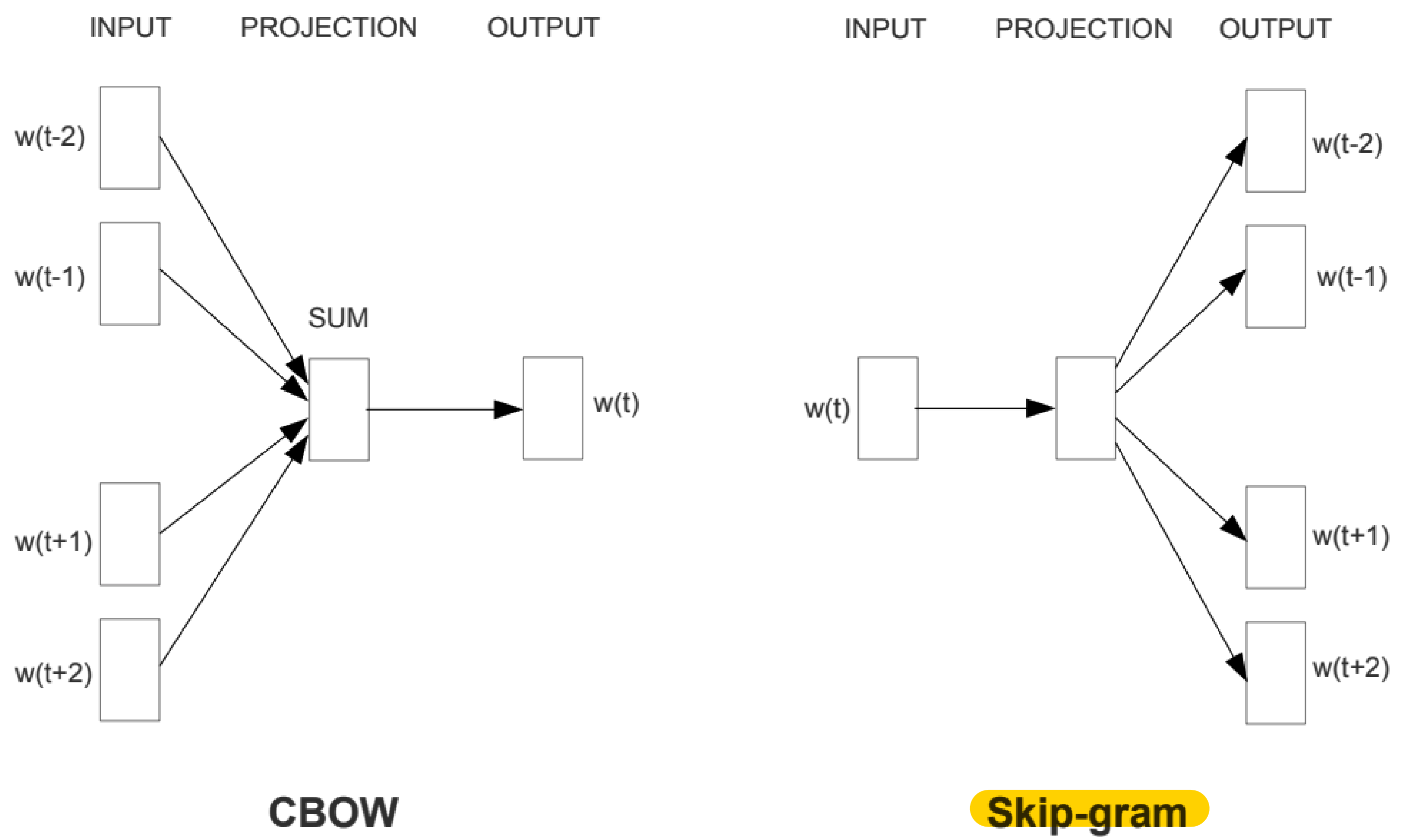
Two Approaches

Count based	Prediction based

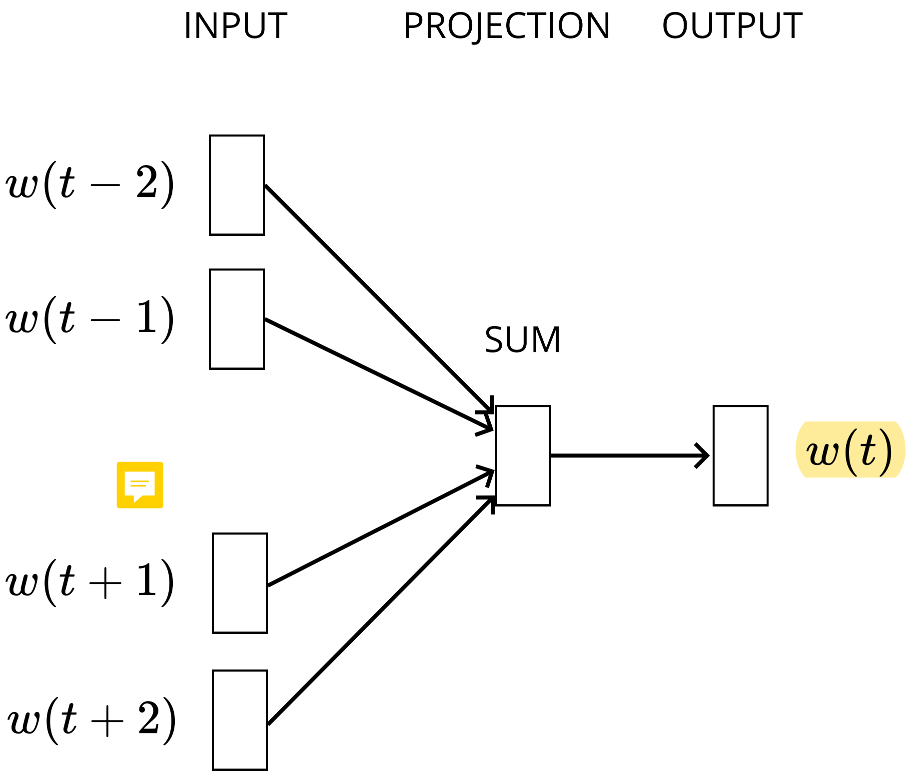
- Directly predict a word from its neighboring words.
- This process will yield dense word embeddings

Word2Vec

Tomas Mikolov et.(2013)



Word2Vec



CBOW

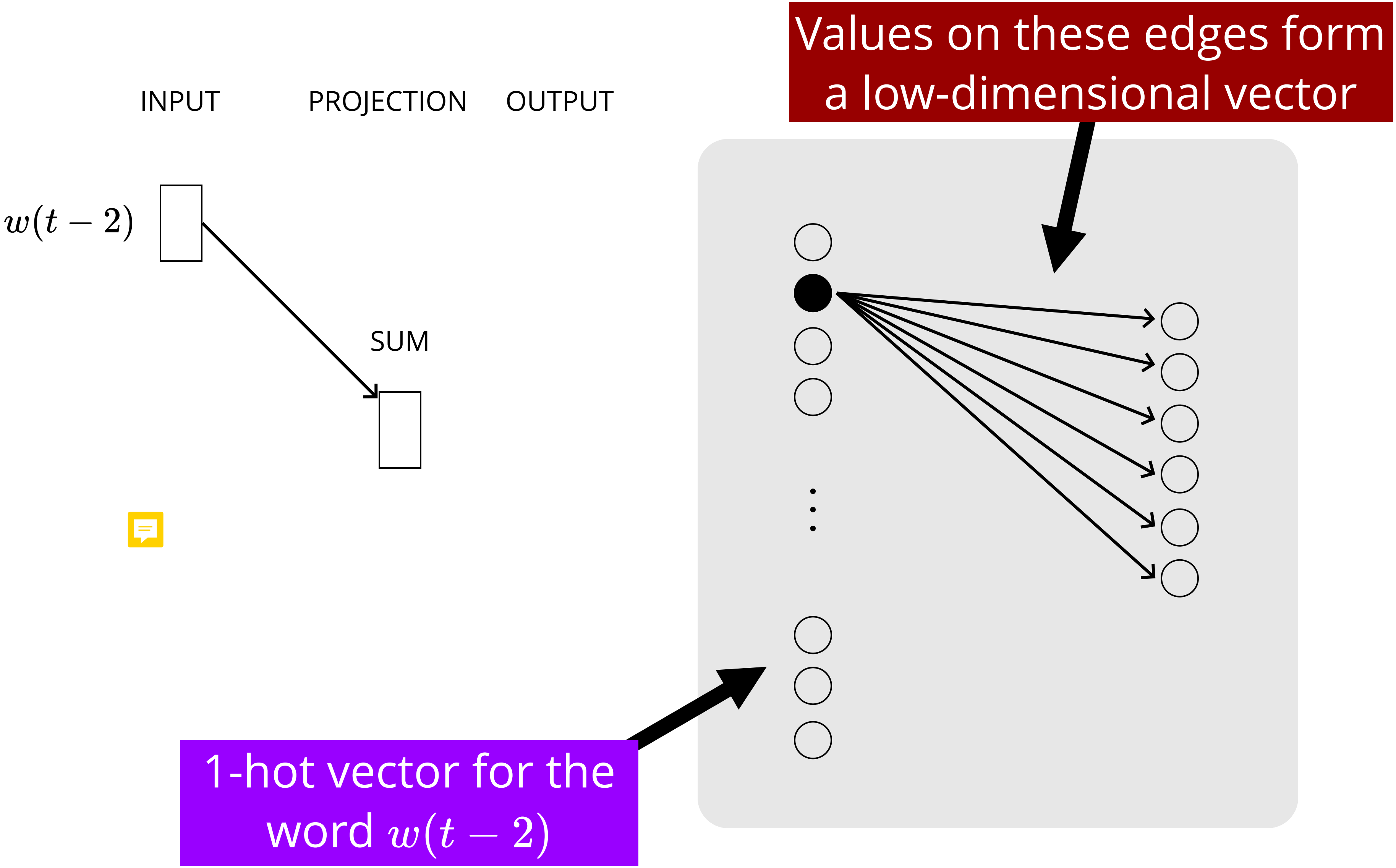
1. Project each context word from $|V|$ dimensional space to its (current) embedding space

2. Sum these vectors

3. Project it to another space of dimension $|V|$ (we expect this vector to be similar to the 1-hot vector of the target word)

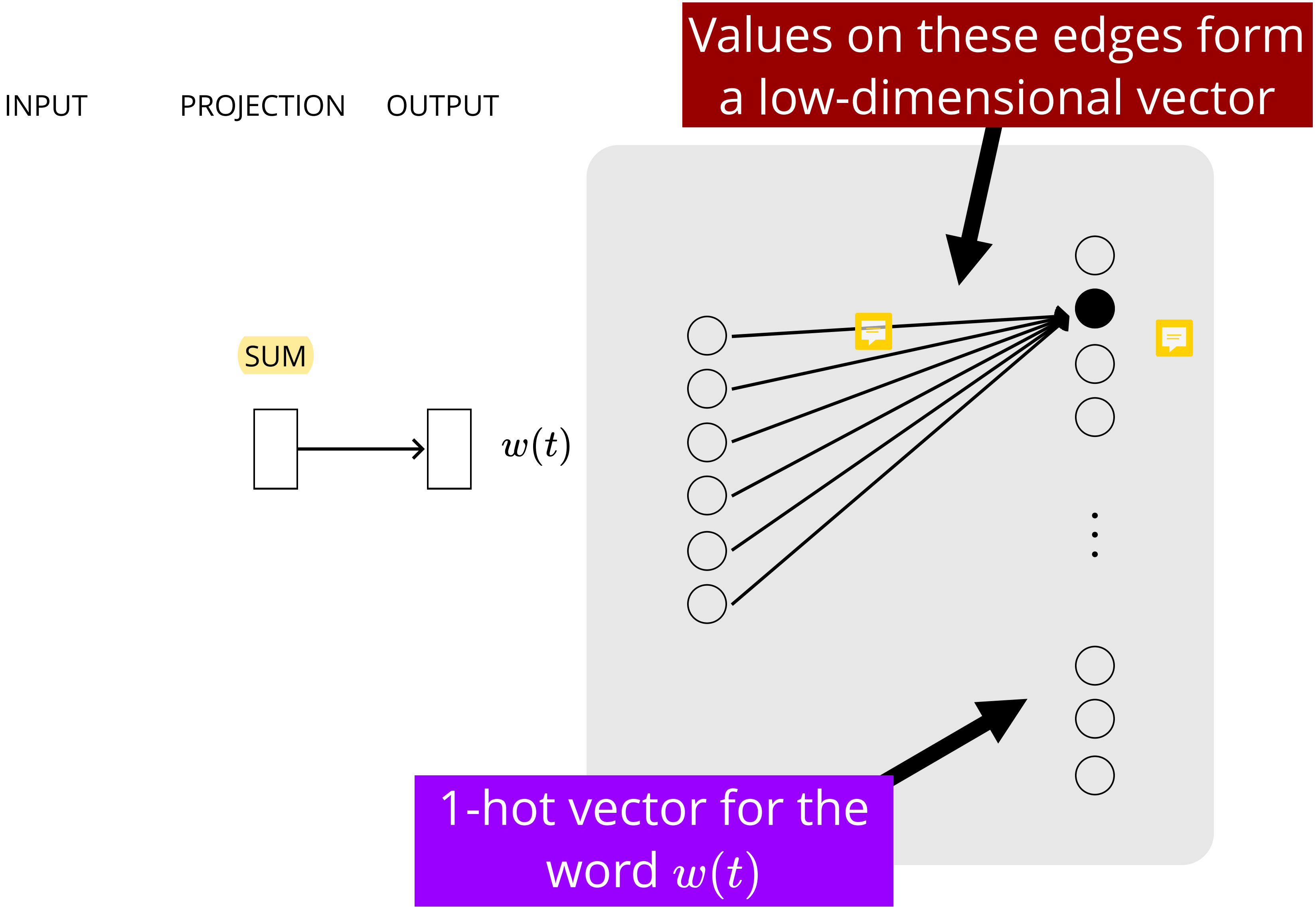
Where are the embeddings?

Word2Vec



INPUT
PROJECTION
OUTPUT
$w(t-2)$
SUM
Values on these edges form a low-dimensional vector
1-hot vector for the word $w(t-2)$

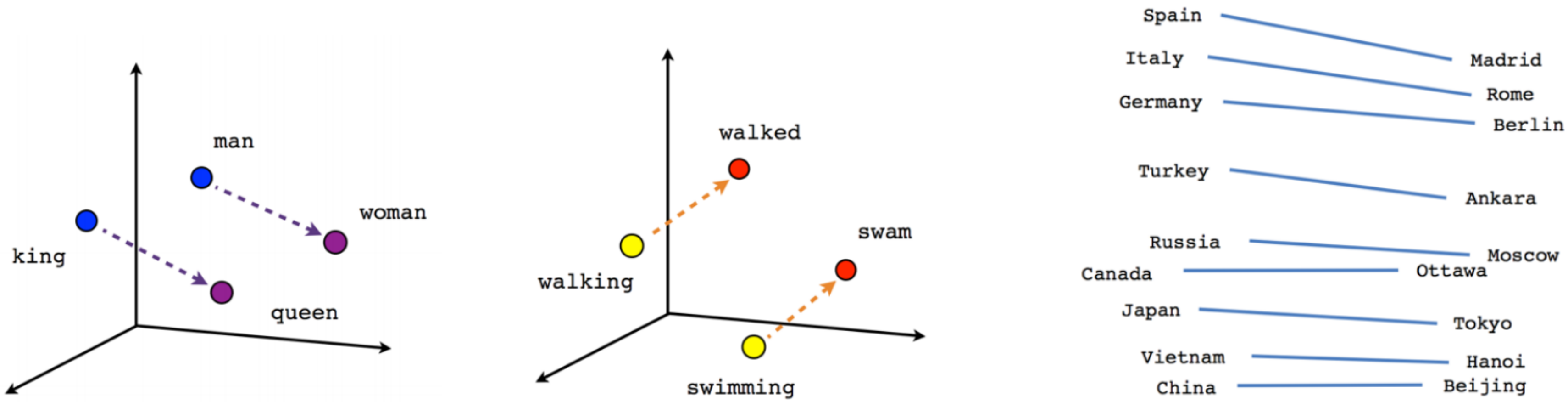
Word2Vec



Values on these edges form a low-dimensional vector
INPUT
PROJECTION
OUTPUT
SUM
$w(t)$
1-hot vector for the word $w(t)$

Word2Vec

Some early interesting properties found in word embeddings

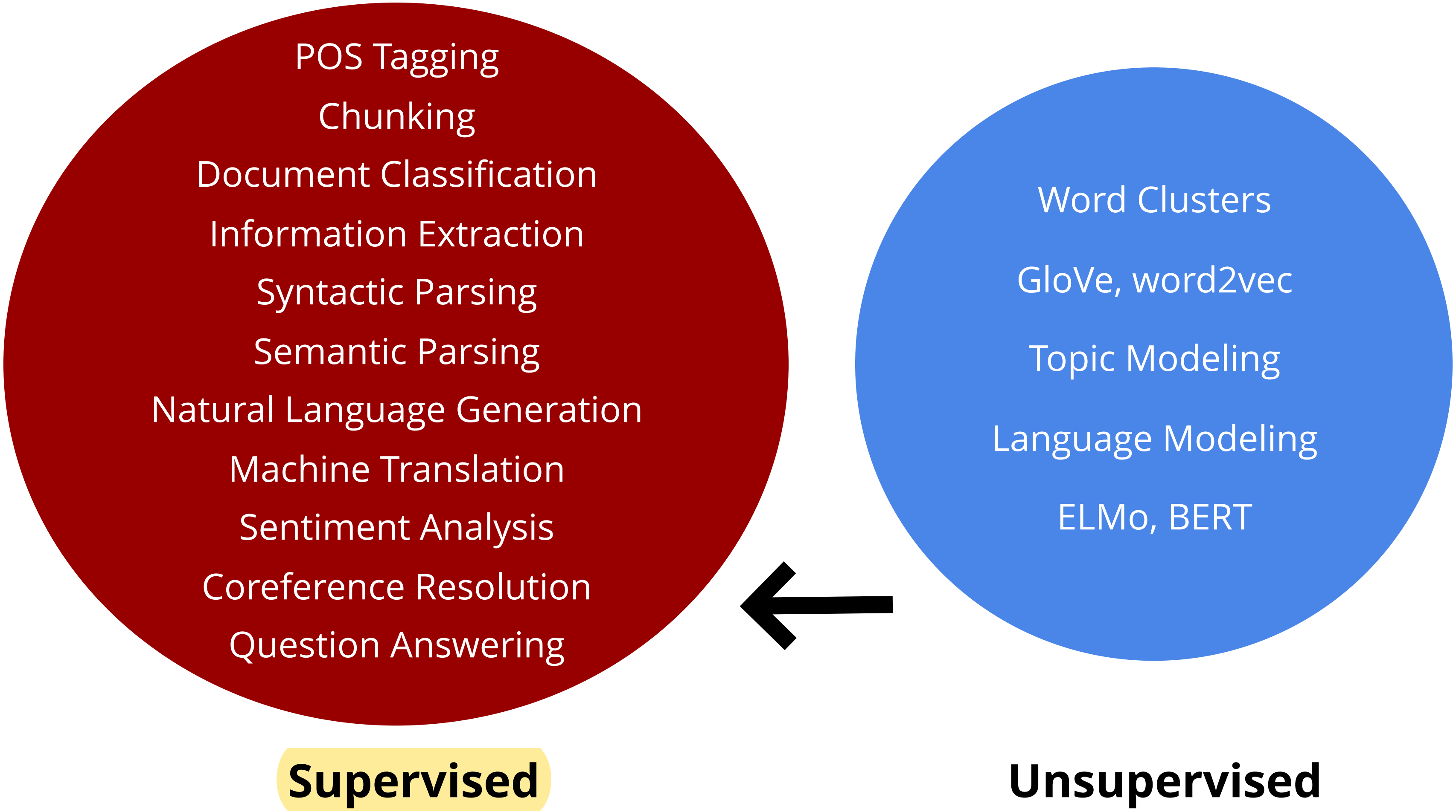


Male-Female

Verb tense

Country-Capital

Tasks in NLP


POS Tagging
Chunking
Document Classification
Information Extraction
Syntactic Parsing
Semantic Parsing
Natural Language Generation
Machine Translation
Sentiment Analysis
Coreference Resolution
Question Answering
Word Clusters
GloVe, word2vec
Topic Modeling
Language Modeling
ELMo, BERT
Supervised
Unsupervised