**Entity Extraction (Domain Specific Unstructured Text)**

**Agenda**

1. Motivation/ Entity Extraction
2. Recurrent Neural Network
3. Word Embedding

**Entity Extraction**

* Subtask of information extraction
* Also known as Named-entity recognition (NER), entity chunking and entity identification
* Find phrases in text that refer to a real-world entity of specific types

For ex,

Mr. Zoran is at Strata in San Jose on March 7.

Zoran: Person

Strata: Organization

San Jose: Location

March 7: Date

Why it is useful?

**Indexing**

* e.g., find all people in a document collection

**Discovery**

* e.g., learn about new drugs from recent biomedical articles

**Relation Extraction**

* Works (Person, Organization), e.g., Works (Zoran, Strata)

**Question Answering**

* Where is Zoran? Zoran is in San Jose

**Customer Entity Extraction**

* Pre-trained models for common entity types (Person, Organization, location, date, etc)
* Custom Models
  + New entity types (e.g., drug, disease)
  + Different language properties
    - Foreign Names
    - Non-proper sentences (e.g., tweets)

**Domain Specific Entity Extraction**

**Biomedical Named Entity Recognition**

* Entity types

Drug/Chemical, disease, protein, DNA, etc,

* Critical Step for complex biomedical NLP task
* Extraction of diseases, Symptoms from electronic medical or health records
* Understanding the interactions between different entity types such as drug-drug interaction, drug-disease relationship and gene-protein relationship, e.g.,
* Drug A cures Disease B
* Drug A Causes Disease B

Similar for other Domains (e.g., legal, finance, Invoice)

**Machine Learning Approach**

* Dictionary approach
* What if a phrase is not in the dictionary

Zoran is at Strata in San Jose on March 7.

Mr. Zoran is at Strata in San Jose on March 7.

Zoran Smith is at Strata in San Jose on March 7.

* Statistical (ML) Approach
* Takes features from surrounding words

Zoran is at Strata in San Jose on March 7.

Mr. Zoran is at Strata in San Jose on March 7.

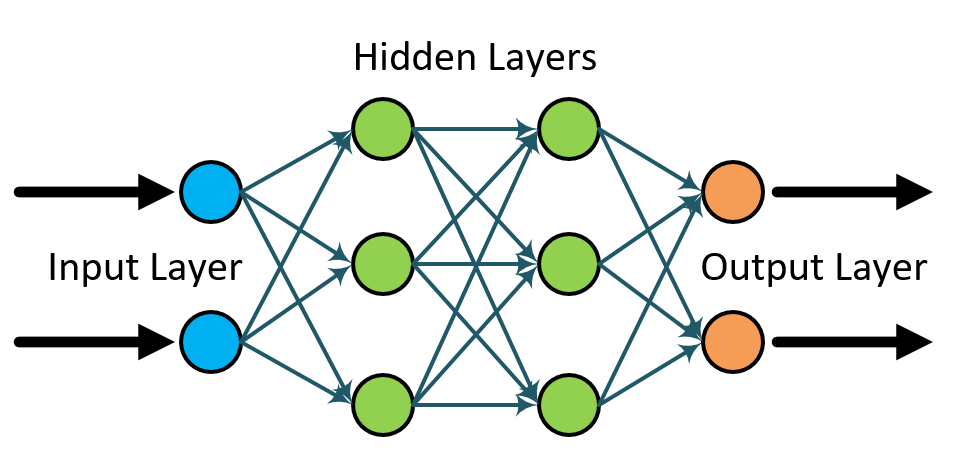
Zoran Smith is at Strata in San Jose on March 7.

Clues: after “Mr.”, before known last name (“Smith”)

* Structured prediction (sequence tagging) – label for a word depends on other labels
* Rather than classifying each word independently.

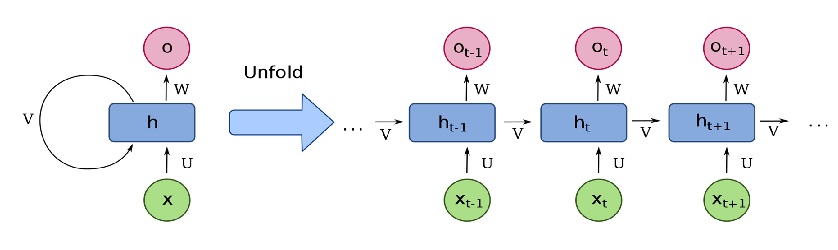
**Deep Learning Approach:**

* Can find complex relationships between input and output using
* Non-linear processing units
* Multiple hidden layers



* Special-purpose layers/architectures have been developed for different problems
* Recurrent Neural Networks (RNNs) are commonly used for sequence-labelling tasks
* Long Short-Term Memory layer (LSTM)
* Comparison to Conditional Random Fields (CRFs)
* CRFs are linear model’s w/o hidden layers
* CRFs have short-term memory

**Recurrent Neural Network**



* The state ht denotes the memory of the network and is responsible for capturing information about previous time steps.
* All layers in RNNs share the same set of parameters because we are performing the same task at each time step with different inputs.

**Word Embedding**

* Features – words
* Features – One-Hot Encoding

Word2Vec

* Simple neural network of a single hidden layer with a linear activation function (Skip-Gram, CBOW)
* Unsupervised learning from large corpora. The word vectors (embedding) are learned by the stochastic gradient descent optimization algorithm
* Publicly available pre-trained models such as Google News
* Can we do better on a specific domain?

