Named Entity Extraction
From Online News

GROUP3
Akshakt Kashyap
Bernardo Najlis
Fady Fadel

Agenda

- Name Entity Extraction
- Problem Definition
- Previous Work and Literature Review
- Dataset (Gold Corpus) and Challenges
- Model.1 / Maximum Entropy Markov Model
- Model.2 / Deep Neural Network with LSTM
- Conclusions

Named Entity Extraction

- “Subtask of Information Extraction that seeks to locate and classify named entities in text into pre-defined categories:
  persons, organizations, locations”
  Definition Source: Wikipedia - https://en.wikipedia.org/wiki/Named-entity_recognition

- Extracts information out of unstructured data (news articles, emails, blog posts, scientific papers)

- *Similar* to Part-of-Speech tagging:
  instead of looking for POS we are looking for entities

- *Not so similar* to Part-of-Speech tagging:
  an entity can span multiple words (i.e.: Terry Fox, Rogers Communications, South Korea)

Problem Definition

- Research has shown that NER systems developed for a specific domain do not perform well against other domains.

- Named Entities are open word class problem
  Basic NER models rely on list of entities **(gazetteer)** to identify them, such list can be expensive to maintain.

- Considerable effort is required in training NER for new domains.

- State of art NER systems rely heavily on **hand crafted features** that works only on certain languages.

- **Our Goal:** Develop machine learning models that predict named entities without any of the above. Compare models performance.

Previous Work / Literature Review

- **Named Entity Recognition using Support Vector Machine**
  https://pdfs.semanticscholar.org/d94a/6a0cd9e03faf6e70814c8053305f01e2c885.pdf
- **Named Entity Recognition with a Maximum Entropy Approach**
  www.comp.nus.edu.sg/~nght/pubs/conll03.pdf
- **Named Entity Recognition using Hidden Markov Model (HMM)**
  https://pdfs.semanticscholar.org/9528/4b31f27b9b8901fdc18554603610ebbc2752.pdf
- **Biomedical named entity recognition using conditional random fields and rich feature sets**
  https://dl.acm.org/citation.cfm?id=1567618
- **Named Entity Recognition with Bidirectional LSTM-CNNs**
  https://www.aclweb.org/anthology/Q16-1026
- **GloVe: Global Vectors for Word Representation**
  https://www.aclweb.org/anthology/D14-1162

Dataset (Gold Corpus) & Challenges

- **Dataset:**
  - Globe and Mail news for period of February 2018 to March 2018
  - 2,116 articles - 70,554 sentences - 1,685,626 unigrams and IOB tags containing 110,032 entities
  - Used **assisted tagging** method to obtain Gold Corpus(IOB Format)
    Raw Data → Preprocess → **SpaCy** → Manual validation → IOB-Format

Justin	Trudeau	visits	Rio	Tinto	aluminum	factory	in	Hamilton
B-PER	I-PER	O	B-ORG	I-ORG	O	O	O	B-GEO

Dataset (Gold Corpus) & Challenges

- **Challenges:**
  - Difficult to build pre-tagged dataset.
  - Found single dataset that applies to our domain.
  - Cleaning, tagging and refining the data takes about 70% of the time.
  - Mapping entities names:
    SpaCy labels → CoNLL labels

SPACY	NLTK
ORG	ORGANIZATION
PERSON	PERSON
LOC	LOCATION
DATE	DATE
TIME	TIME
MONEY	MONEY
PERCENT	PERCENT
FACILITY	FACILITY
GPE	GPE

Model.1 / Maximum Entropy Markov Model

- In his famous 1957 paper, Ed. T. Jaynes wrote:
  - *Information theory provides a constructive criterion for setting up probability distributions on the basis of partial knowledge, and leads to a type of statistical inference which is called the maximum entropy estimate*.

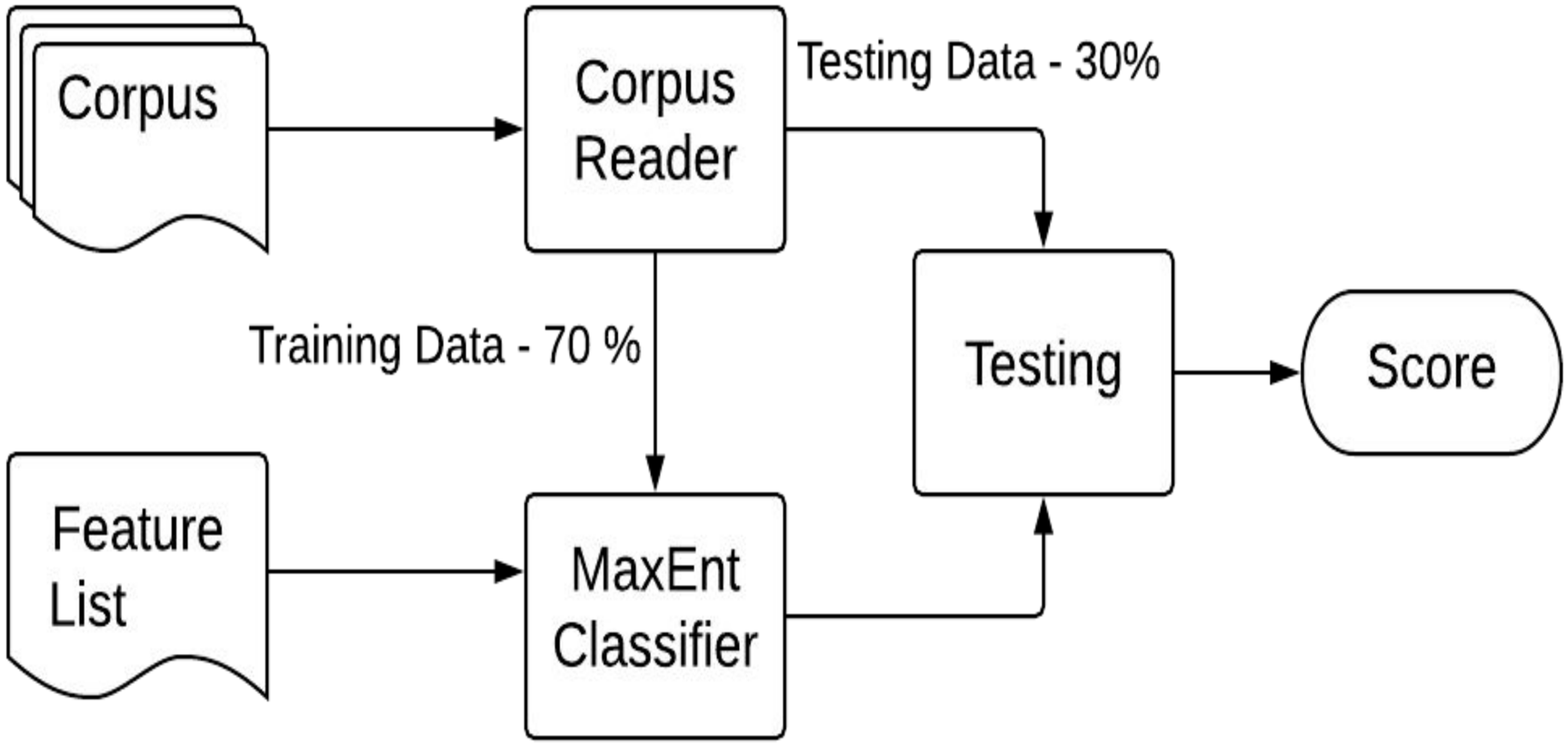
  - ***It is least biased estimate possible on the given information****; i.e., it is maximally noncommittal with regard to missing information.*

- *Maximum Entropy was first introduced to NLP area by Berger, et al (1996) and Della Pietra, et al. 1997. Since then, Maximum Entropy technique (and the more general framework Random Fields) has enjoyed intensive research in NLP community*.

Model.1 / Maximum Entropy Markov Model

- **Steps:**
  - Read the data using CoNLLCorpusReader.
  - Use NLTK MaxEnt classifier with MEGAM. MEGAM (MEGA Model Optimization Package) is an OCaml based Maximum Entropy project that originated from Utah university. MEGAM tends to perform much better in terms of speed and resource consumption.
  - Create feature list and feed it to the MaxEnt Classifier:
    - Current Word
    - Current POS
    - Next Word / NextNext Word
    - Next POS / NextNext POS
    - Prev Word / PrevPrev POS
    - Prev IOB
    - Surrounding POS tag sequence
    - Capitalized words
    - POS tag sequence after “DT” tag

Model.1 / Maximum Entropy Markov Model



Model.1 / Results

- **Results:**
  - Obtained 93.8% accuracy using 70% training / 30% testing.

- **Limitations:**

  - New entities require additional features for training the classifier.

  - Accuracy plateaus at a certain number of features.

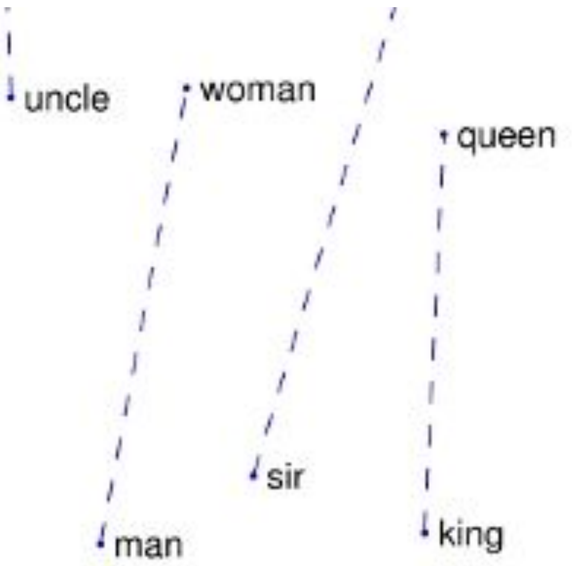
  - Features have different contribution rates.

Model.2 / Deep Neural Net with LSTM

- **Steps:**
  - Read data using NLTK CoNLL Corpus Reader
  - Encode and pad words to max sentence length
  - One-hot encode and pad labels
  - Load GloVe embeddings
  - Layers:
    - Input units: as many as max sentence length
    - Output units: as many as IOB tags / labels

    Embedding → LSTM / Bidirectional LSTM → Dropout → Sigmoid

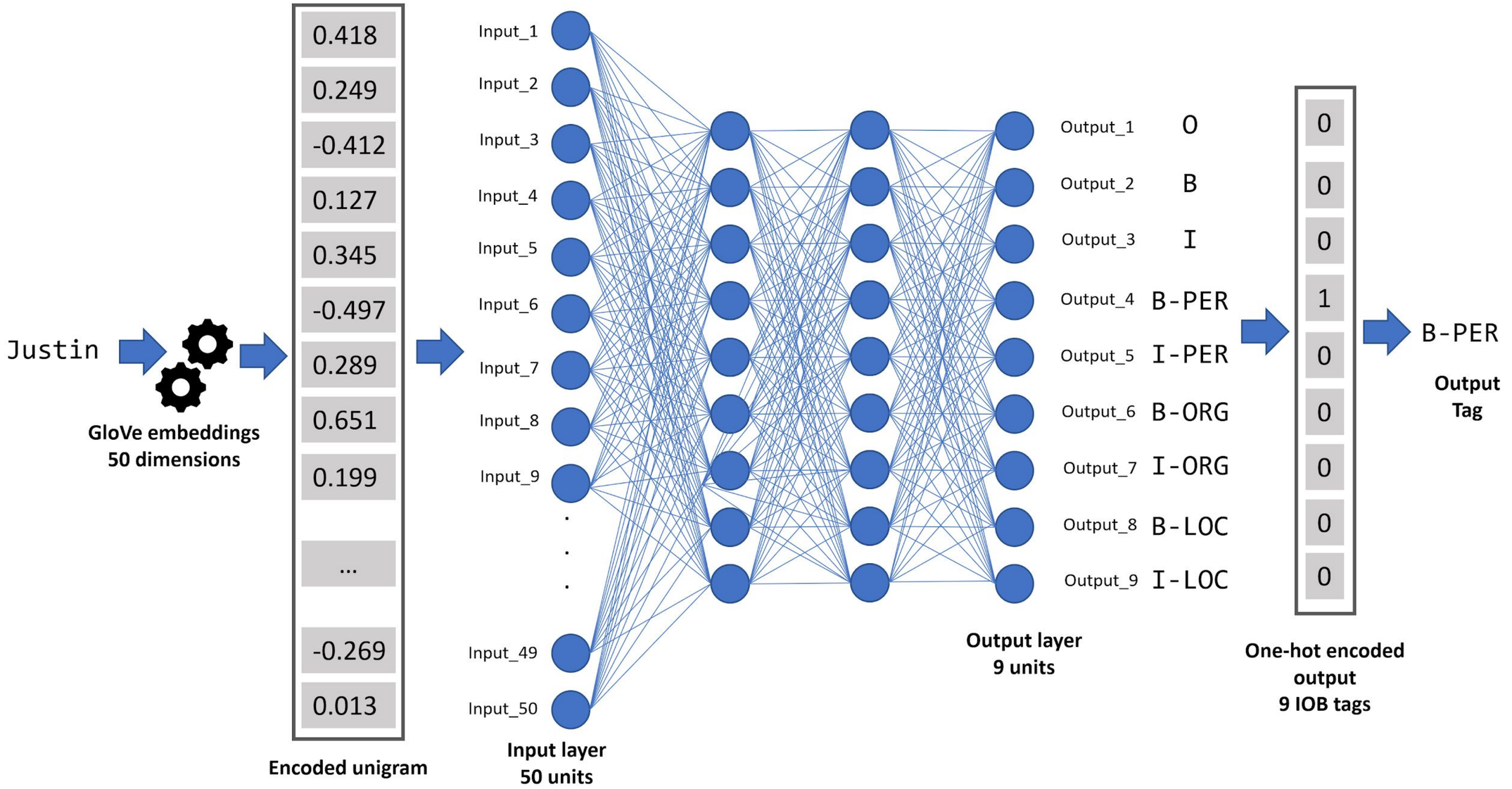
Why Word Embeddings?



- Word Embeddings map words to vectors of real numbers (Embedding from a space with one dimension per word to a continuous vector space)
- Similar unigrams with similar semantics have similar directions
- GloVe (Global Vectors) is developed as an open-source project at Stanford
- Calculated using ratios of word-word co-occurrence probabilities
- Faster to calculate for large corpus, outperforms CBOW and SkipGrams

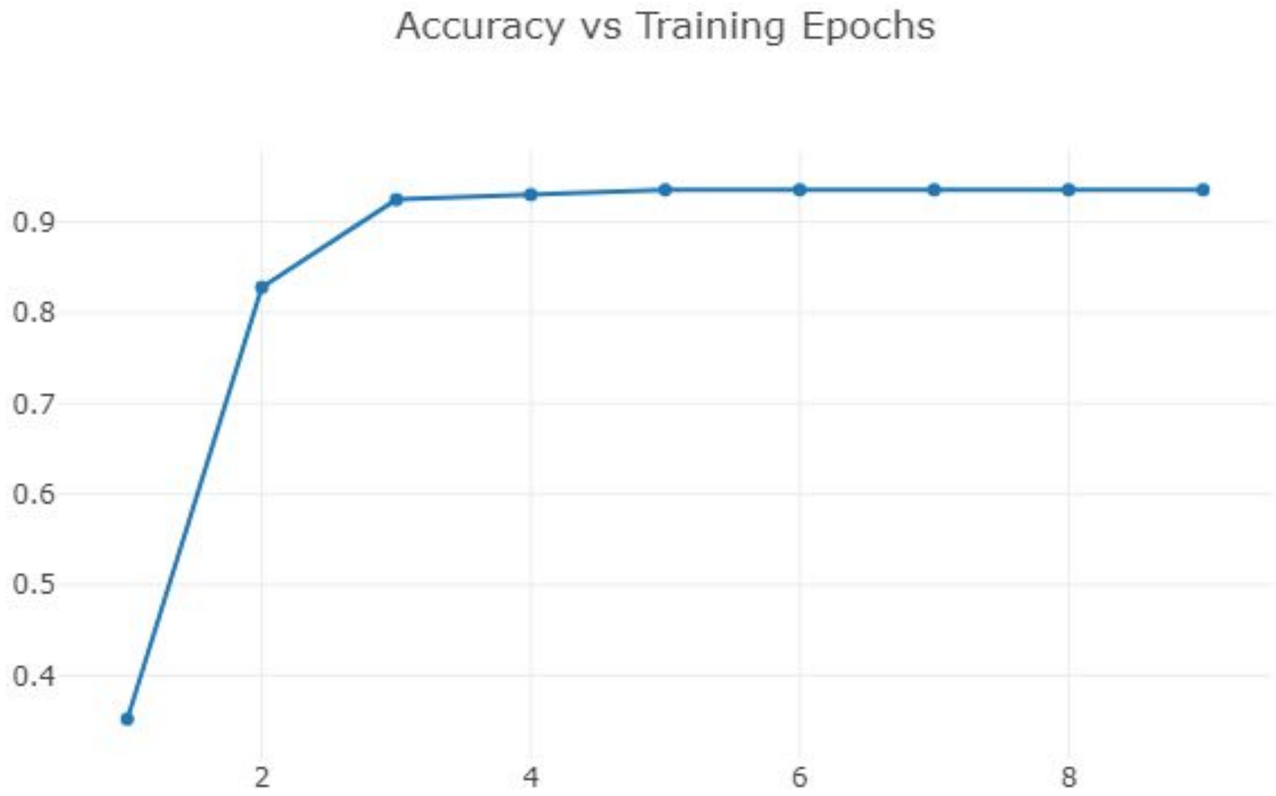
Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Model.2 / Deep Neural Net with LSTM



Model.2 / DNN with LSTM

- **Results:**
  - Tested GloVe embeddings with 50, 100, 200 and 300 dimensions
  - Tested 1 to 10 epochs, small and large dataset
  - Obtained **93.5% accuracy** using **70% training** / **30 % testing**
  - Accuracy plateaus after **6 epochs**
- **Limitations:**
  - Most of time spent on encoding and padding data



Conclusions

- Named Entity Gold dataset is difficult to obtain (manual effort required).
- MaxEnt accuracy (93.8%) and DNN accuracy (93.5%) are comparable
- MaxEnt performance is limited by the added features / DNN is dependent on the embeddings available (pretrained vs custom)
- MaxEnt requires domain knowledge to increase feature accuracy / DNN does not require domain knowledge
- MaxEnt would perform poorly on untrained domain / DNN is more generalizable
- MaxEnt requires less computing resources / DNNs require more compute power for model training on regular size datasets

Workload Distribution

TASK	A.K.	B.N.	F.F.
Data Collection	0%	0%	100%
Data preparation	0%	0%	100%
Data Validation	20%	20%	60%
Literature Review	35%	35%	30%
Model Feasibility Assessment	40%	40%	20%
Maxent Model + Features	80%	10%	10%
Deep Neural Net + Embedding	10%	80%	10%

References and Resources

- Named Entity Recognition using Support Vector
- NLTK MEGAM Max Ent algorithms on Windows
- http://www.nltk.org/book/ch07.html
- http://nlpforhackers.io/named-entity-extraction/
- https://machinelearningmastery.com/use-word-embedding-layers-deep-learning-keras/
- http://legacydirs.umiacs.umd.edu/~hal/megam/
- https://homepages.inf.ed.ac.uk/lzhang10/maxent.html
- http://www.cs.cmu.edu/afs/cs/user/aberger/www/html/tutorial/node3.html