A Gentle Introduction

Natural Language Processing

Why
Natural Language Processing?

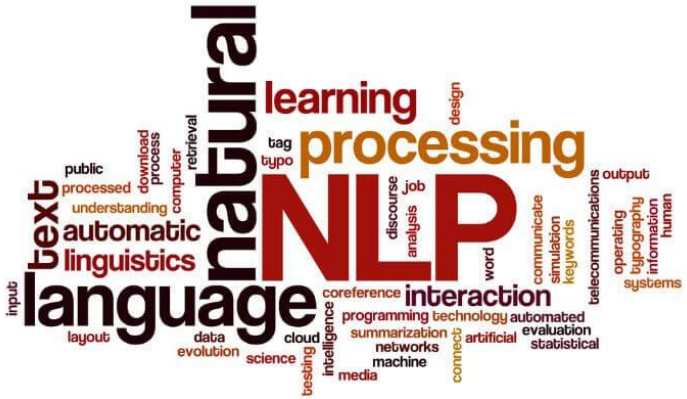
- Computing devices have become a huge part of our lives
- we generate tonnes and tonnes of data
- we desire to know what the data is telling us
- sometimes in real-time

Why
Natural Language Processing?

- We need our devices to “speak” to us
- We desire systems that can listen to us
- Understand what we say in human language
- Act on our instructions and …
- Communicate to us in human language

So what is Natural Language Processing?

- a set of tools and techniques for processing human languages
- enables effective human-machine interaction
- extracts information from some generated data (often unstructured)



So what is Natural Language Processing?

- defines how information is represented
- parsing that information from the data generating process
- construct and use efficient data structures that stores the model

Definition of Natural Language Processing?

In summary NLP is often defined as:

- the study of the computational treatment of natural (human) language

It aims to:

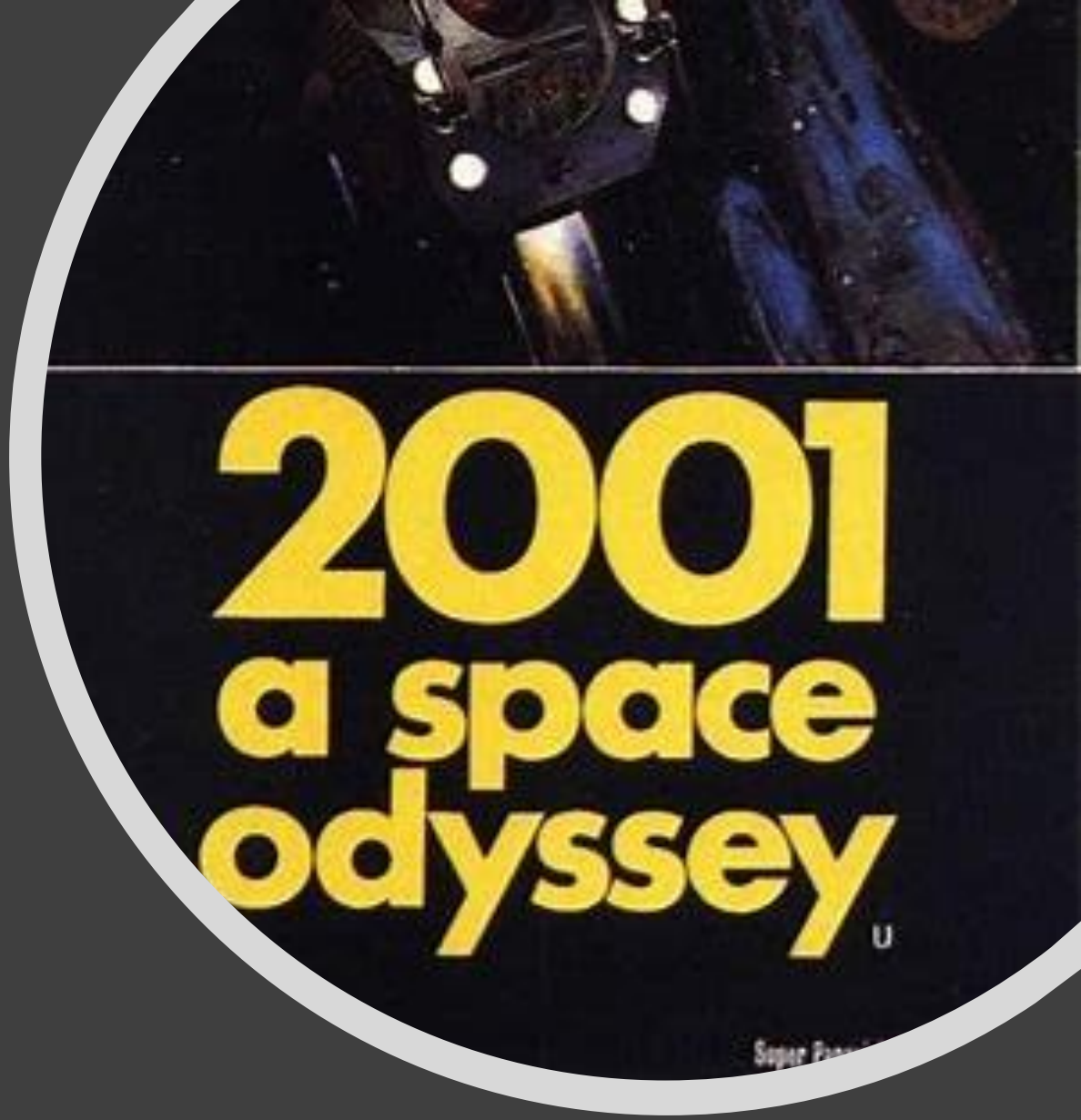
- create systems that *understand* and *generate* (produce) human language

A Classic NLP Example?

- Quote from the 1968 SciFi movie "2001: A Space Odyssey" by *Stanley Kubrick*

  Dave Bowman: *Open the pod bay doors, HAL.*

  HAL: *I'm sorry Dave. I'm afraid I can't do that*



Basic NLP Acronyms

- NLP (Natural Language Processing)
- CL (Computational Linguistics)
- IR (Information Retrieval)
- IE (Information Extraction)
- HLT (Human Language Technology)
- NLE (Natural Language Engineering)
- ML (Machine Learning)

Common Applications of NLP

Machine Translation

MT deals with the study of the use of computer systems to translate text or speech from one language to another

Information Retrieval

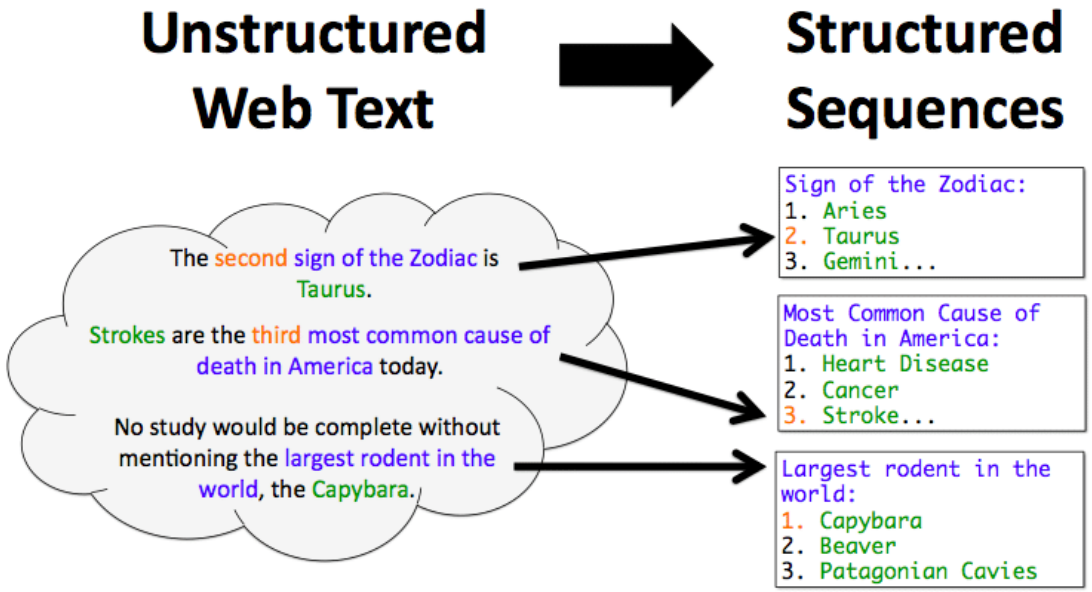
obtaining relevant materials based on
a query from a collection of
information system resources

Sentiment Analysis aka *opinion mining*

refers to the use of natural language processing techniques and tools to identify, extract, quantify, and study affective states and subjective information.

Information Extraction

automatically extraction of structured information from unstructured and/or semi-structured machine-readable documents.



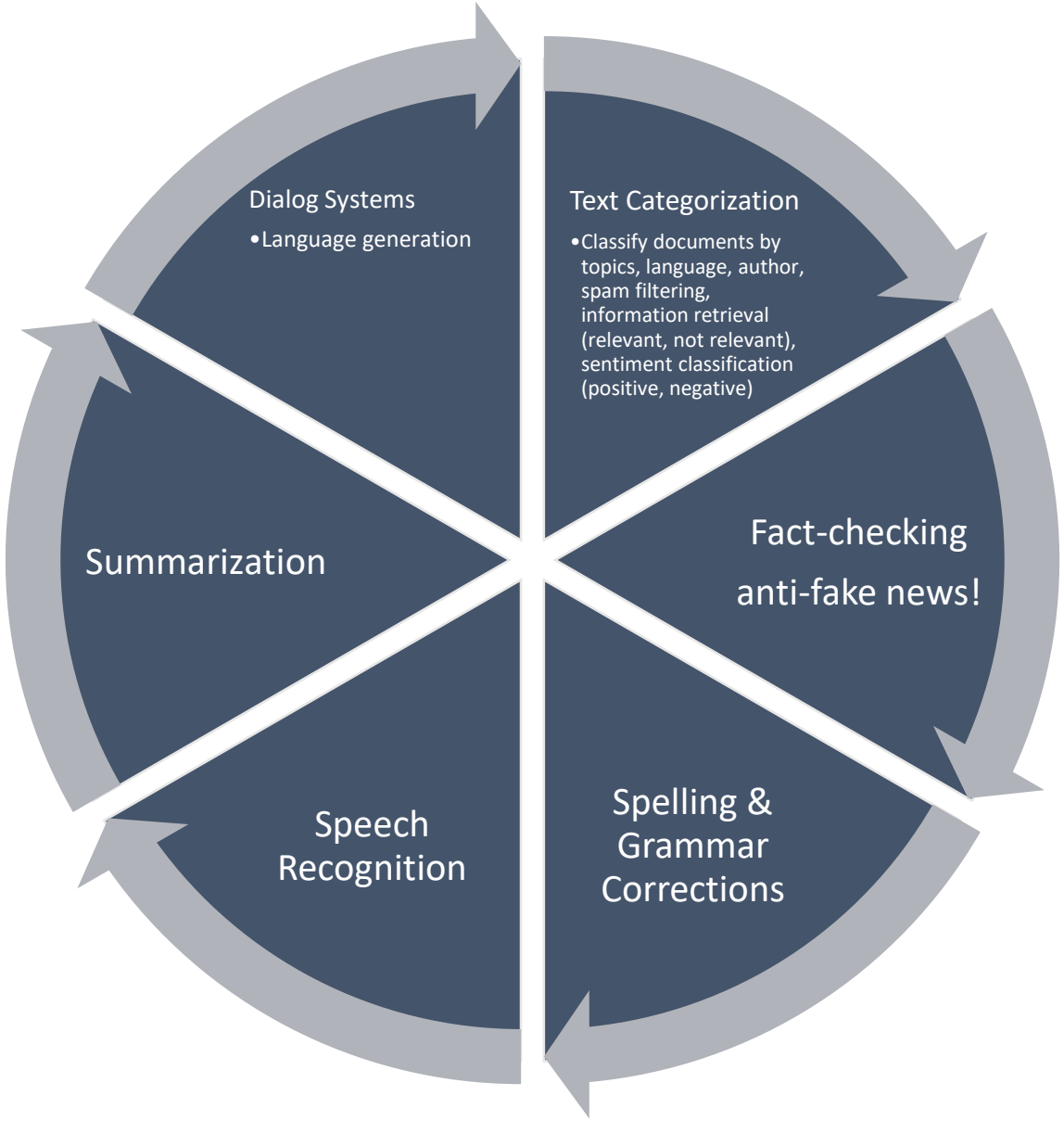
Question Answering

QA systems that answer questions posed by humans with natural language

IBM Watson®

InventionMachine

Other Related NLP Applications



Why is NLP hard?

Humans Language Ambiguity

lexical, phrase, semantic ambiguities

- Iraqi Head Seeks Arms
  - *Word sense is ambiguous (head, arms)*
- Stolen Painting Found by Tree
  - *Thematic role is ambiguous: tree is agent or location?*
- Ban on Nude Dancing on Governor's Desk
  - *Syntactic structure (attachment) is ambiguous: is the ban or the dancing on the desk?*
- Hospitals Are Sued by 7 Foot Doctors
  - *Semantics is ambiguous : what is 7 foot?*

Language evolves

- new words are constantly introduced
- the parsing rules are flexible
- ambiguity is inherent
- meanings are context dependent

Language is subtle

- He arrived at the lecture
- He chuckled at the lecture
- He chuckled his way through the lecture
- **He arrived his way through the lecture
- Language is complicated!

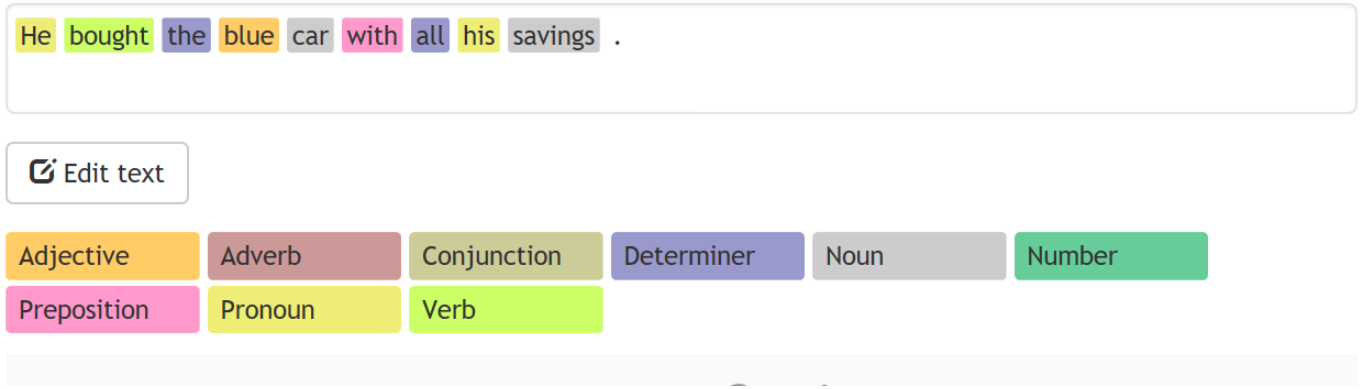
Language representation is unique

- no know universal representation
- often application-specific
- world knowledge required for interpretation
- many languages, dialects, styles etc.

Some key NLP tasks?

Part of Speech Tagging

- Parts of speech tagging is one of the most fundamental tasks in NLP.
- It involves using some techniques and statistics to understand the part of speech of a word.



Parsing

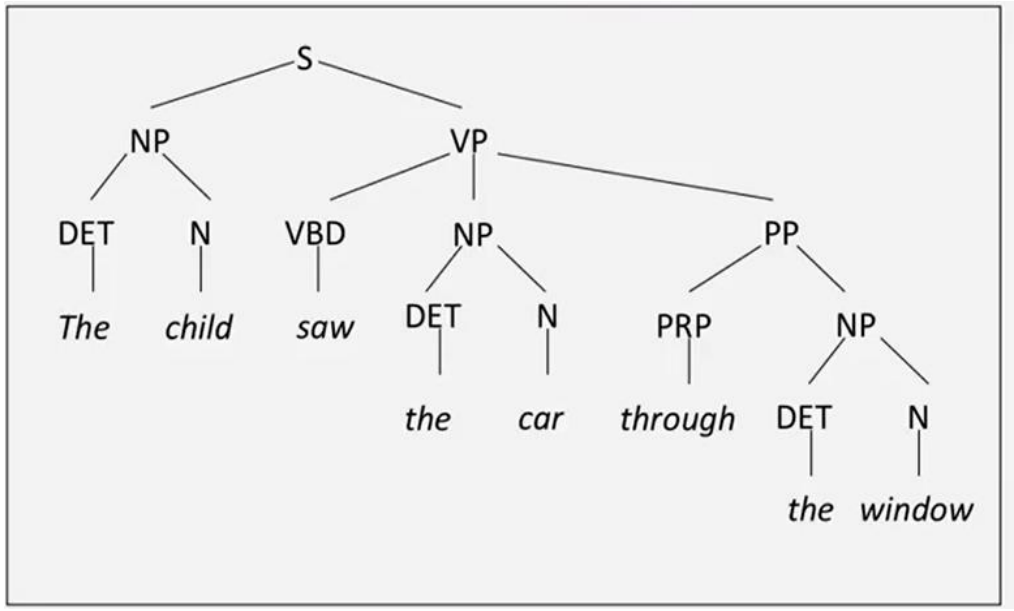
- Parsing attempts to the build the *parse tree* of a sentence based on its grammar
- A grammar could be *phrase-structured* or *dependency grammar*

Phrase-Structure Grammar

```
S   → NP  VP
NP  → DET N
NP  → NP  PP
VP  → VBD
VP  → VBD NP
VP  → VBD NP NP
VP  → VP  PP
PP  → PRP NP
```

```
DET → the
DET → that
DET → a
N   → child
N   → window
N   → car
VBD → found
VBD → ate
VBD → saw
PRP → in
PRP → of
PRP → through
```

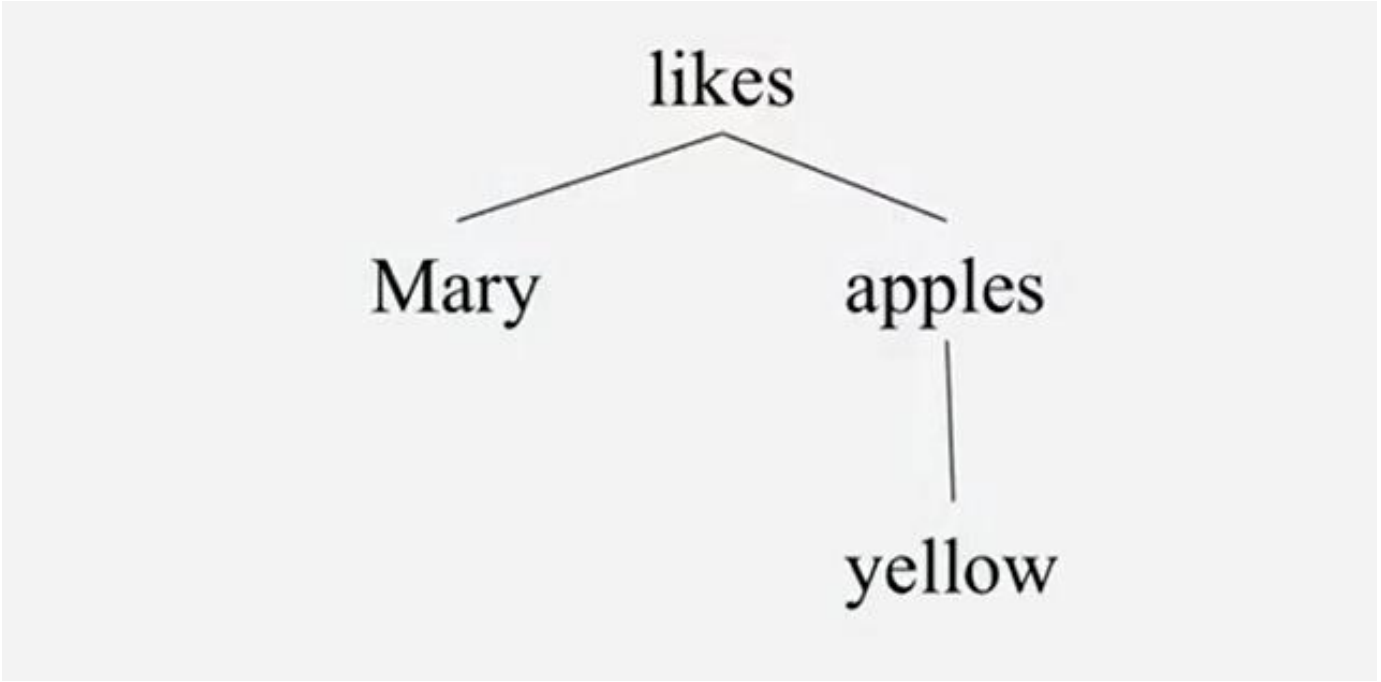
Parse Tree



Parsing

- Parsing attempts to the build the *parse tree* of a sentence based on its grammar
- A grammar could be *phrase-structured* or *dependency grammar*

Dependency Parse Tree for
Mary likes yellow apple



Semantic Analysis

- Semantics analysis is a very essential task in NLP.
- It deals with the logical representation of sentences in forms like first other logic
- Inferences could easily be drawn from sum representations

$$\forall x, y: Mother(x, y) \Longrightarrow Parent(x, y)$$

Semantic Role Labelling

- Semantic role labelling assigns multiple arguments to verbs which may or may not be required
- From the example, the main verb, “accept-V” has other arguments as labelled
- “A3: attribute” was not required in this case

He wouldn’t accept anything of value from those he was writing about.

Semantic Role Labelling

- Semantic role labelling assigns multiple arguments to verbs which may or may not be required
- From the example, the main verb, “accept-V” has other arguments as labelled
- “A3: attribute” was not required in this case

He wouldn’t <u>accept</u> anything of value from those he was writing about.

[$_{\textbf{A0}}$ *He*] [$_{\textbf{AM-MOD}}$ *would*] [$_{\textbf{AM-NEG}}$ *n’t*] [$_{\textbf{V}}$ **accept**]
[$_{\textbf{A1}}$ *anything of value*] from [$_{\textbf{A2}}$ *those he was writing about*]

Semantic Role Labelling

- Semantic role labelling assigns multiple arguments to verbs which may or may not be required
- From the example, the main verb, “accept-V” has other arguments as labelled
- “A3: attribute” was not required in this case

He wouldn’t <u>accept</u> anything of value from those he was writing about.

[$_{A0}$ *He*] [$_{AM-MOD}$ *would*] [$_{AM-NEG}$ *n’t*] [$_{V}$ **accept**] [$_{A1}$ *anything of value*] from [$_{A2}$ *those he was writing about*]

- **V:** verb
- **A0:** acceptor
- **A1:** thing accepted
- **A2:** accepted-from
- **A3:** attribute
- **AM-MOD:** modal
- **AM-NEG:** negation

Word Sense Disambiguation

- Word sense disambiguation resolves sense ambiguities
- It uses the context in which the word appears in to determine which sense is more appropriate
- Sense inventories such as WordNet or electronic dictionaries are often used
- Useful in Machine Translation

Typically aims at resolving sense ambiguities e.g. ***bank***

The fisherman jumped off the **bank** into the water.

The **bank** down the street was robbed!

We **bank** on his ability to deliver on the task.

The **bank** in that road is too steep and dangerous

Word Sense Disambiguation

- Word sense disambiguation resolves sense ambiguities
- It uses the context in which the word appears in to determine which sense is more appropriate
- Sense inventories such as WordNet or electronic dictionaries are often used
- Useful in Machine Translation

Typically aims at resolving sense ambiguities e.g. ***bank***

The fisherman jumped off the **bank$_1$** into the water.

The **bank$_2$** down the street was robbed!

We **bank$_3$** on his ability to deliver on the task.

The **bank$_4$** in that road is too steep and dangerous

Named Entity Recognition

- NER identifies the named entities in a sentence such as persons, locations and organisation
- Where there is a combinations of words forming a named entity:
- the first label starts with “B”: e.g. B-PERS
- The inside labels starts with “I” i.e. I-LOC
- Often used in information extraction

Wolff, currently a journalist in Argentina, played with Del Bosque in the final years of the seventies in Real Madrid.

http://cogcomp.cs.illinois.edu/page/demo_view/NER
http://nlp.stanford.edu:8080/ner/

Wolff	B-PER
,	O
currently	O
a	O
journalist	O
in	O
Argentina	B-LOC
,	O
played	O
with	O
Del	B-PER
Bosque	I-PER
in	O
the	O
final	O
years	O
of	O
the	O
seventies	O
in	O
Real	B-ORG
Madrid	I-ORG
.	O

Coreference Resolution

- Tries to understand when two phrases are referring to one person.
- This is often used in discuss to avoid repetition
- Anaphoric: entity mentioned first and then referred to later
- Cataphoric: entity referred to before being mentioned

- Barack Obama visited China. The US president met with his Chinese counterpart.
- Cynthia went to see her aunt at the hospital. She was scheduled for surgery on Monday.
- Because he was sick, Michael stayed home on Friday.

Basic Language Models

I don't know ____ to go out or not . : weather/whether
We went ____ the door to get inside . : through/threw
They all had a ____ of the cake . : piece/peace
She had to go to ____ to prove she was innocent . : caught/court
We were only ____ to visit at certain times . : aloud/allowed
She went back to ____ she had locked the door . : cheque/check
Can you ____ me ? : hear/here
Do you usually eat ____ for breakfast ? : serial/cereal
She normally ____ with her mouth closed . : chews/choose
I'm going to ____ it on the internet . : cell/sell

Consider this simple problem

Any idea how to get the computer to choose the right word to fill the blank?

Probabilistic Language Models

Which of the following is more probable?

- *Time flies like an arrow.*
- *Fruit flies like a banana.*

Probabilistic Language Models

- Language models define the probability distribution over sequence of words (or sentences)
- That is:
  - *How likely is a given sentence to appear in the language?*

Probabilistic Language Models

- Common applications:
  - Speech recognition
    - P(“recognize speech”) > P(“recognize speech”)
  - Text Generation
    - P(“three houses”) > P(“three house”)
  - Spelling correction
    - P(“my cat eats fish”) > P(“my xat eats fish”)
  - Auto Completion
    - P(“give ___”[me]) > P(“give ___”[i])
  - Machine Translation
    - P(“the blue house”) > P(“the house blue”)
  - Other uses:
    - OCR
    - Summarization
    - Document classification

Probabilistic Language Models

Given a sentence with n, words: $S = [w_1, w_2, w_3, \dots, w_n]$ the language model computes the probability:

- $P(S_n) = P(w_1, w_2, w_3, \dots, w_n)$

Predicting the next word w_{n+1}:

- The probability of S_{n+1}, $P(S_{n+1})$
- $P(w_{n+1}|w_1, w_2, w_3, \dots, w_n)$
- *?*

Probabilistic Language Models

- Predicting the next word w_{n+1}:
  - The probability of S_{n+1}, $P(S_{n+1})$
  - $P(w_{n+1}|w_1, w_2, w_3, \dots, w_n)$
- Using the chain rule:
  - $P(w_1)P(w_2|w_1) \dots P(w_{n+1}|w_1, w_2 \dots w_n)$
- Example:
  $P(\text{“}I\ love\ Naija\ jollof\ rice\text{”})$?

Probabilistic Language Models

- Example:
  - $P(\text{“}I\ love\ Naija\ jollof\ rice\text{”})$?
- Solution:
  - $P(\text{“}I\text{”})\ *$
  - $P(\text{“}love\text{”}|\text{“}I\text{”})\ *$
  - $P(\text{“}Naija\text{”}|\text{“}I\text{”},\text{“}love\text{”})\ *$
  - $P(\text{“}jollof\text{”}|\text{“}I\text{”},\text{“}love\text{”},\text{“}Naija\text{”})\ *$
  - $P(\text{“}rice\text{”}|\text{“}I\text{”},\text{“}love\text{”},\text{“}Naija\text{”},\text{“}jollof\text{”})$

Probabilistic Language Models

- So predicting the next word in:
  - "I love Naija jollof ____" [yam/rice]
- Compare:
  - $P(\text{“}yam\text{”}|\text{“}I\text{”},\text{“}love\text{”},\text{“}Naija\text{”},\text{“}jollof\text{”})$

  and
  - $P(\text{“}rice\text{”}|\text{“}I\text{”},\text{“}love\text{”},\text{“}Naija\text{”},\text{“}jollof\text{”})$

N-Gram Models

- Text = “I love Naija jollof rice”
- 1-gram *aka* unigram: “I”, “love”, “Naija”, “jollof”, “rice”
- 2-gram *aka* bigram: “I love”, “love Naija”, “Naija jollof”, “jollof rice”
- 3-gram *aka* trigram: “I love Naija”, “love Naija jollof”, “Naija jollof rice”

Bigram Counts

- Words that follow “your”?
- norvig.com/ngrams/count_2w.txt

What word follows “your”?

- http://norvig.com/ngrams/count_2w.txt

your abilities	160848
your ability	1116122
your ablum	112926
your academic	274761
your acceptance	783544
your access	492555
your accommodation	320408
your account	8149940
your accounting	128409
your accounts	257118
your action	121057
your actions	492448
your activation	459379
your active	140797
your activities	226183
your activity	156213
your actual	302488
your ad	1450485
your address	1611337
your admin	117943
your ads	264771
your advantage	242238
your adventure	109658
your advert	101178
your advertisement	172783

N-Gram Models

- Markov Assumption:
  - Only use a limited history..
- So given: $S = \text{“}I\ love\ Naija\ jollof\ rice\text{”}, P(S)$ is
  unigram: $P(\text{“}I\text{"})P(\text{“}love\text{”})P(\text{“}Naija\text{”})P(\text{“}jollof\text{”})P(\text{“}rice\text{”})$
  bigram: $P(\text{“}I)$P(“love”|“I) … $P(\text{“}rice\text{”}|\text{“}jollof\text{”})$
  trigram:$P(\text{“}I\text{"})P(\text{“}love\text{”}|\text{“}I\text{"})P(\text{“}Naija\text{”}|\text{“}I\text{"},\text{“}love\text{”}))$ …

Estimating N-Gram Probabilities

- Maximum likelihood estimation:

$$P(w_1|w_2) = \frac{count(w_1,w_2)}{count(w_1)}$$

- Counts collected over a large (hundreds of billions) text *corpus*[1] over available on the web

[1]*Corpus* (pl. corpora) is simply a collection of text

I don't know ____ to go out or not . : weather/whether
We went ____ the door to get inside . : through/threw
They all had a ____ of the cake . : piece/peace
She had to go to ____ to prove she was innocent . : caught/court
We were only ____ to visit at certain times . : aloud/allowed
She went back to ____ she had locked the door . : cheque/check
Can you ____ me ? : hear/here
Do you usually eat ____ for breakfast ? : serial/cereal
She normally ____ with her mouth closed . : chews/choose
I'm going to ____ it on the internet . : cell/sell

Looking at this again

Any idea how to get the computer to choose the right word to fill the blank?

N-Gram Models

- “I don’t know _____ to go out or not”
- options: “weather” vs “whether”
- unigram: $P(\text{“}weather\text{”})$ vs $P(\text{“}whether\text{”})$
- bigram: $P(\text{“}weather\text{”}|\text{"know"})$ vs $P(\text{“}whether\text{”}|\text{"know"})$
- trigram:
  - $P(\text{“}weather\text{”}|\text{"don't know"})$ vs $P(\text{“}whether\text{”}|\text{"don't know"})$

Data Sparsity and Smoothing

- We may not see all the counts we need in our text corpus
- What if a word (say “Naija”) is not in the text corpus

$$P(Naija|I\ love) = 0$$

because:

$$count(Naija) = 0$$

- We use a technique called *smoothing* to avoid it

Laplace (Add-One) Smoothing

- For all possible n-grams, add the count of one

$$p = \frac{c + 1}{n + v}$$

where:

- c = count of n-grams in corpus
- n = count of history
- v = vocabulary size

- Add-α: a variant of this technique that adds α to c.

Back-off

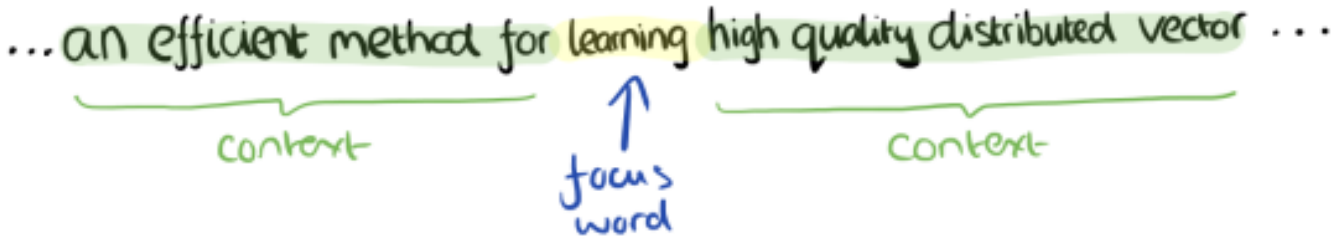
- Back of is also a technique for dealing with unseen n-grams
- With a trigram model, if $count(Naija) = 0$ then:
  - $P(yam|Naija\ jollof)$ and
  - $P(rice|Naija\ jollof)$
- But we still can compare by backing-off to bigram:
  - $P(yam|jollof)$ and
  - $P(rice|jollof)$

Common Smoothing Techniques

- Additive smoothing
- Good-Turing estimate
- Jelinek-Mercer smoothing (interpolation)
- Katz smoothing (backoff)
- Witten-Bell smoothing
- Absolute discounting
- Kneser-Ney smoothing
- A good description of each could be found here:
  - https://nlp.stanford.edu/~wcmac/papers/20050421-smoothing-tutorial.pdf

Word Embedding Models

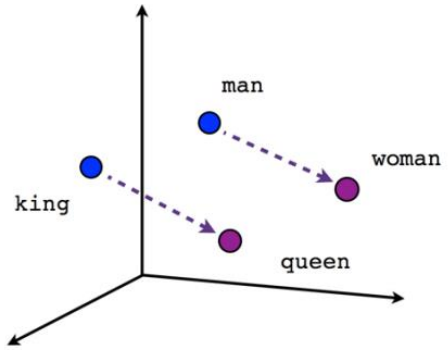
Word embedding models



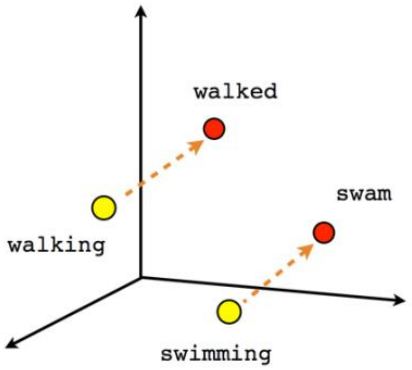
- Been around for a while but made popular by the ***Mikolov et al, 2013***
- *Vector space* models that captures relationships between words surprisingly well
- Uses dense vector representation
- Efficient unsupervised training on data
- Examples:
  - ***word2vec***
  - ***GLoVe***

Word embedding models

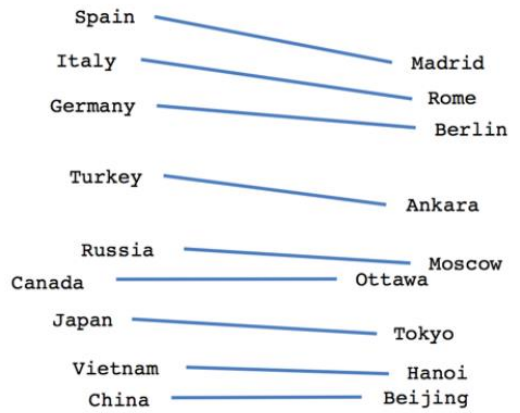
- Embeddings capture interesting relationship from data



Male-Female



Verb tense



Country-Capital

Evaluating embedding models

- Analogical reasoning: It can answer ***woman*** to the question…

$$king : man :: queen : ?$$

- Word similarity: It knows that the closest words to *king* are:

man, queen, prince, castle, etc

- Odd word out:

breakfast, lunch, ***cereal***, dinner

Open Problems in NLP

References and Resources

Books

- Speech and Language Processing
  - Dan Jurafsky and James Martin
  - http://www.cs.Colorado.edu/~martin/slp.html
- Foundations of Statistical Natural Language Processing
  - Chris Manning and Hinrich Schutze
  - https://nlp.Stanford.edu/fsnlp
- Natural Language Understanding
  - James Allen

References and Resources

University Online Courses:

- Johns Hopkins University (Jason Eisner)
- Cornell University (Lillian Lee)
- Stanford University (Chris Manning)
- U. Maryland (Hal Daume)
- Berkeley (Dan Klein)
- U. Texas (Ray Mooney)

References and Resources

Other Online Courses:

- Coursera
  - Chris Manning & Dan Jurafsky (2012) Basic Intro
  - Michael Collins (2013, more advance)
- Stanford CoreNLP
- NLTK: Natural Language Toolkit
- Chris Manning (Stanford University) https://youtu.be/OQQ-W_63UgQ
- Dragomir Radev (University of Michigan) https://youtu.be/n25JjoixM3I

References and Resources

Lecture Materials

- Andreas Vlachos: (Cambridge University)
  - http://andreasvlachos.github.io//teaching/
- Mark Hepple: (University of Sheffield)
  - https://staffwww.dcs.shef.ac.uk/people/M.Hepple/campus_only/COM3110/

References and Resources

- Major Conferences (and Journals):
  - ACL = Annual Meeting of the Association of Computational Linguistics
  - NAACL-HLT = Annual Conf North American Association of Computational Linguistics
  - EMNLP = Conference on Empirical Methods in Natural Language Processing
  - COLING = International Conference on Computational Linguistics
  - NIPS -
  - EACL = European Association of Computational Linguistics
  - LREC = Language Resources and Evaluation Conference
  - INTERSPEECH = Conference of the International Speech Communication Association

References and Resources

Books

- Speech and Language Processing
  - Dan Jurafsky and James Martin
  - http://www.cs.Colorado.edu/~martin/slp.html
- Foundations of Statistical Natural Language Processing
  - Chris Manning and Hinrich Schutze
  - https://nlp.Stanford.edu/fsnlp
- Natural Language Understanding
  - James Allen

For listening…