Machine Learning for Vision and Language

MSc Artificial Intelligence

Lecturer: Tejaswini Deoskar
t.deoskar@uu.nl

UiL-OTS, Utrecht University

Block 1, 2019-20

some slides and examples adapted from S. Goldwater, A. Louis

What does it mean to process language?

“I’ve had a wonderful weekend!

I always wanted to buy a melodica.
On Saturday, I finally went to that fancy music store in Haarlem.
The rest of the weekend, I practised some of my favourite songs on it.
On Monday ..."

image credit: https://commons.wikimedia.org/

People infer a range of things from text or speech

- Meaning of words, phrases, and sentences
- Resolve co-referring expressions
  - ∗ “it” refers to the melodica
- Relationships between sentences
  - ∗ I went *because* I wanted to buy a melodica
- Implied information
  - ∗ I bought the melodica at that store in Haarlem
  - ∗ The new melodica is the one I practised songs on
- Have impressions about the speaker/writing style.
  - ∗ The writing is boring or funny or engaging..

All of this understanding plays a role when we

- Make conversations with other
- Translate from one language to another
- Find an answer to a question from a text
- Create a summary of a document
- ...

- Natural Language Processing/Computational Linguistics in AI:
  - How to study/analyse language in computational terms?
  - How can we do some of these tasks algorithmically/artificially?

Goals of NLP/CL

- Scientific Goal: Build models of the human use of language
  - ∗ Humans are better at language processing than machines
  - ∗ Language is complex; modelling can help
- Technological Goal: Build models that serve in technological applications
  - ∗ machine translation, speech systems, information extraction, etc.

Can be far-reaching, or relatively modest

- True text/speech understanding, reasoning/decision-making from text
- more intelligent web-search, spelling-correction in context.

What are the basic tasks a language processing system should solve

- Uncover the basic levels of structure in language

Words

This is a simple sentence WORDS

Morphology

The study of sub-word units of meaning.

This	is	a	simple	sentence	**WORDS**

	`be`
	`3sg`
	`present`
					MORPHOLOGY

Parts of Speech

Classes of words (VERB, NOUN, ADJECTIVE, PREPOSITION etc.)

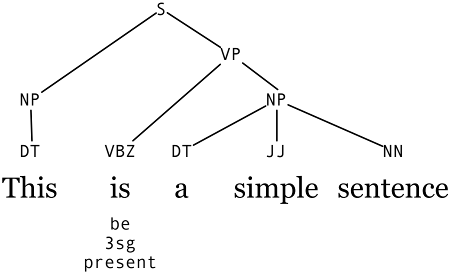
DT	VBZ	DT	JJ	NN	**PART OF SPEECH**
This	is	a	simple	sentence	**WORDS**
	be 3sg present				**MORPHOLOGY**

Knowledge of Parts of Speech and Morphology

Knowledge of word classes and morphology is part of our linguistic knowledge (independent of semantics).

Those zorls you splarded were malgy.

Syntax



SYNTAX

PART OF SPEECH

WORDS

MORPHOLOGY

Syntax

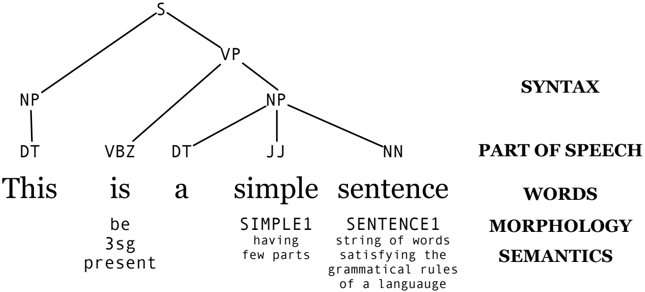
- Study of the structural relationships between words and phrases.
- Words are organised into groups (called phrases) which function as a unit.

Dogs sleep soundly
My next-door neighbours sleep soundly
Green ideas sleep soundly

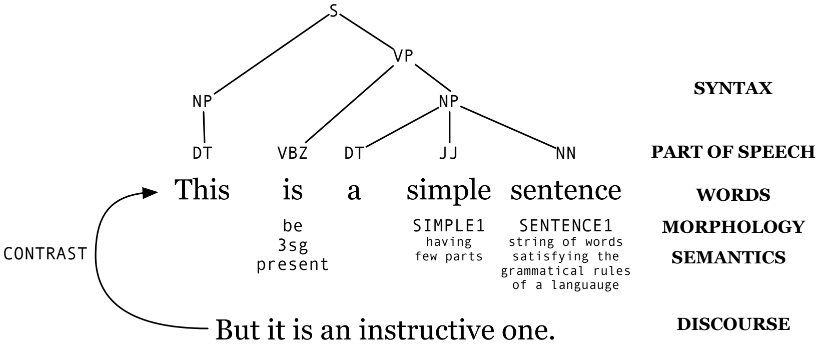
- *I [like bananas] but [hate pineapples].* [VP- co-ordination]
- *I like bananas. Do you?* [VP-ellipses]

Semantics

Lexical Semantics and Compositional Semantics



Discourse



S

VP

NP NP SYNTAX

DT VBZ DT JJ NN PART OF SPEECH

This is a simple sentence WORDS

be
3sg
present

SIMPLE1
having few parts

SENTENCE1
string of words satisfying the grammatical rules of a languauge

MORPHOLOGY

SEMANTICS

CONTRAST

But it is an instructive one. DISCOURSE

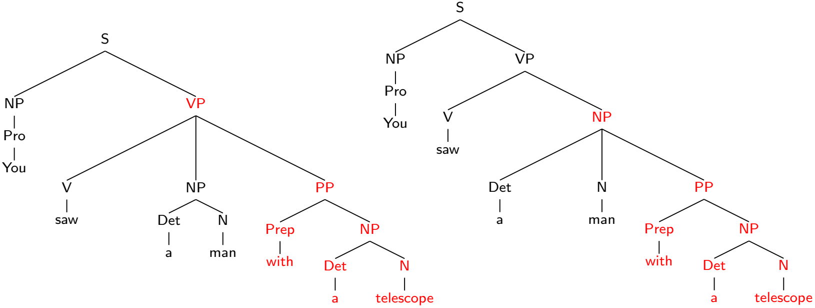
Why is ML for Language hard?

Language is **ambiguous** at many levels.

- Word level:
  - ∗ Most words have multiple meanings (senses): bank (financial institute or river?)
  - ∗ Many words have "vague" meaning: a small elephant > a big rabbit
- Parts of speech: chair (noun or verb?), box, can, ...
- Syntactic structure: You saw a man with a telescope

Ambiguity grows with sentence length, often in exponential manner.

Syntactic Ambiguity



Ambiguity grows with sentence length, often in exponential manner.

Semantic Ambiguity

- Quantifiers: *some, all, many, a, one,* etc.

- Quantifier scope: *Every* boy gave a flower to *some* girl.

  The company sent a new battery to every car owner.

  The company sent a doctor to treat every employee affected by the accident.

Humans can disambiguate very well, machines not so well.

Why is ML for Language hard? Variability

Did Google buy YouTube?

1. Google purchased YouTube
2. Google's acquisition of YouTube
3. Google acquired every company
4. YouTube may be sold to Google
5. Google didn't take over YouTube

Example from "Combined Distributional and Logical Semantics", Lewis & Steedman, TACL 2013

Why is ML for Language hard? Context dependence and Unknown representations

- correct interpretation is context-dependent and often requires world knowledge.
- Very difficult to capture, since we don't know how to represent the knowledge a human has/needs:
  - ∗ What is the "meaning" of a word or sentence?
  - ∗ How to model context?
  - ∗ How to model general knowledge?

Statistical and neural models

Today, language processing is dominated by statistical and/or neural approaches.

- Typically more robust than earlier rule-based methods
- Relevant models and model statistics/probabilities/weights are *learned from data*
  - ∗ Supervised Models
  - ∗ A few cases of "unsupervised" learning (learning from raw data/text)
- Normally requires *lots of data* about any particular phenomenon.

But how much data is enough?

- Natural language data follows a particular distribution captured by Zipf's Law (a power law distribution)
  - ∗ e.g., a few words are very common, but a large majority of words are rare or very infrequent in data.
- There are always linguistic phenomena which are important but have rare evidence in data.

Word Counts

Most frequent words in the English Europarl corpus (out of 24m words (tokens))

any word

Frequency	Token
1,698,599	the
849,256	of
793,731	to
640,257	and
508,560	in
407,638	that
400,467	is
394,778	a
263,040	I

nouns

Frequency	Token
124,598	European
104,325	Mr
92,195	Commission
66,781	President
62,867	Parliament
57,804	Union
53,683	report
53,547	Council
45,842	States

Word Counts

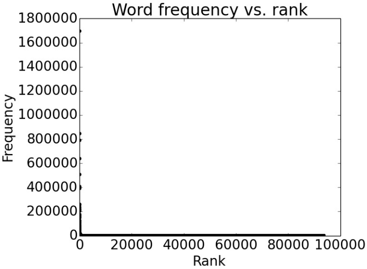
Out of 93638 distinct words (word types), 36231 occur *only once*.

Examples:

- cornflakes, mathematicians, fuzziness, jumbling
- pseudo-rapporteur, lobby-ridden, perfunctorily,
- Lycketoft, UNCITRAL, H-0695
- policyfor, Commissioneris, 145.95, 27a

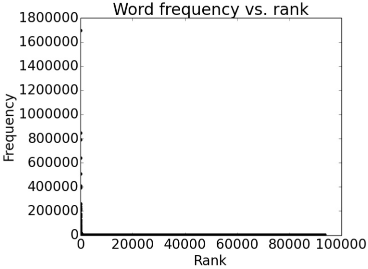
Plotting word frequencies

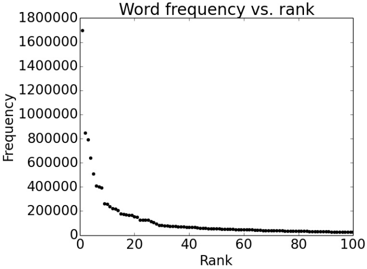
- Order words by frequency.
- 1st rank is the highest-frequency word, 2nd rank is the second most frequent word, etc.



Plotting word frequencies

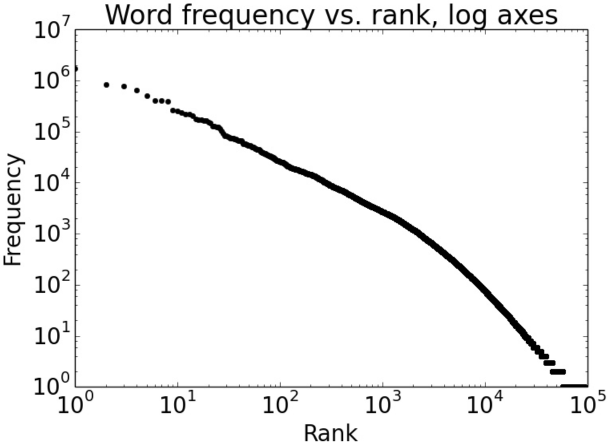
Order words by frequency. What is the frequency of nth ranked word?





Rescaling the axes

To really see what's going on, use logarithmic axes:



Zipf's law

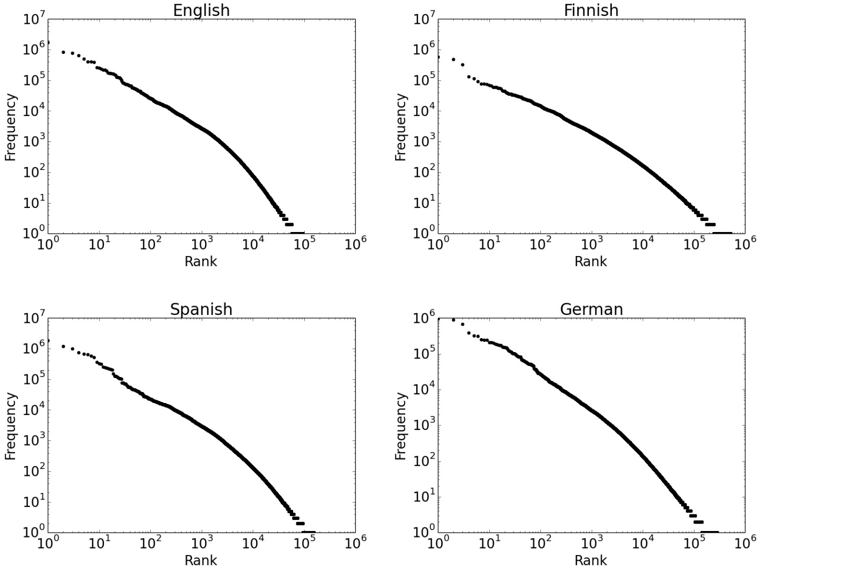
Summarizes the behaviour we just saw:

$$f \times r \approx k$$

- f = frequency of a word
- r = rank of a word (if sorted by frequency)
- k = a constant

- Why a line in log-scales?
- $fr = k \quad \Rightarrow \quad f = \frac{k}{r} \quad \Rightarrow \quad \log f = \log k - \log r$

What about other languages?



Implications of Zipf's Law for data-driven/ML/neural models

- Regardless of how large our corpus is, there will be a lot of infrequent (and zero-frequency!) words.
- In fact, the same holds for many other levels of linguistic structure (e.g., syntactic rules).
- This means our models should have good generalisation ability.

Language processing and Deep Learning

Michael Jordan: `Although current deep learning research tends to claim to encompass NLP, I'm much less convinced about the strength of the results, compared to the results in, say, vision, and less convinced that the way to go is to couple huge amounts of data with black-box learning architectures.`

Language processing and Deep Learning

Chris Manning: So far, problems in higher-level language processing have not seen the dramatic error rate reductions from deep learning that have been seen in speech recognition and in object recognition in vision. Although there have been gains from deep learning approaches, they have been more modest than sudden 25% or 50% error reductions.

Where has Deep Learning helped NLP? The gains so far have not so much been from true Deep Learning (use of a hierarchy of more abstract representations to promote generalization) as from the use of distributed word representations - through the use of real-valued vector representations of words and concepts.

Part II: Organisation

- Lexical semantics (word-level semantics)
  - ∗ Distributional and Vector based neural models (word embeddings)
- Modelling sequences (RNN based)
  - ∗ Language modelling
  - ∗ Part-of-speech tagging
  - ∗ Supertagging

Part II: Organisation

Assignments are in **Jupyter Notebook**. Follow installation instructions on BB.

- Assignment 2A (ungraded) : A tutorial on pytorch
  - ∗ PyTorch is a python-based deep learning library
  - ∗ increasingly adapted in industry and research
- Assignment 2B (15%): Build word embeddings (lexical semantics)
  - ∗ based on GloVe embeddings , Pennington et al. (2014)
- Assignment 2C (15%): Build a POS tagger/supertagger (RNN based)

All have separate intermediate deadlines (check Blackboard)

Language Models

- Broad sense: models of human language
- Narrow sense (as used in NLP):
  - ∗ probabilitistic models that assign a probability to a sentence/sequence of words
  - ∗ "Probability of a sentence" = how likely is it to occur in natural language
    - ▶ Consider only a specific language (English)

P(She studies morphosyntax) > P(She studies more faux syntax)

P(She is going home) > P(She is going house)

Machine translation

Sentence probabilities help decide word choice and word order.

non-English input

↓ (Translation model)

possible outputs

She is going home
She is going house
She is traveling to home
To home she is going
...

↓ (Language model)

best-guess output

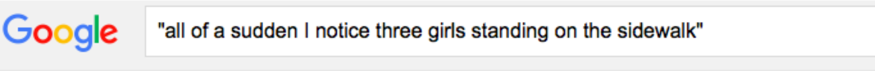
She is going home

Data Scarcity

What do you think will be the estimated probability :

$$P(\text{all of a sudden I notice three girls standing on the sidewalk})$$

Even if I give you the web as a corpus



"all of a sudden I notice three girls standing on the sidewalk"

$$P(\text{all of a sudden I notice three guys standing on the sidewalk}) = 0$$

Data sparsity is a problem

- Most sentences will not be seen even in very large corpora
- We need to decompose the probability

Chain Rule:

$$P(w_1^n) = P(w_1)\ P(w_2|w_1)\ P(w_3|w_1, w_2) \ldots P(w_n|w_1 \ldots w_{n-1})$$

We still have to estimate probabilities for long sequences from corpus data.

Make an approximation and keep only limited history

$$P(w_1^n) = P(w_1)P(w_2|w_1)P(w_3|w_1,w_2)\ldots P(w_n|w_1\ldots w_{n-1})$$

- History of size 2

  $P(w_1^n) \approx P(w_1)P(w_2|w_1)P(w_3|w_1,w_2)\ldots P(w_n|w_{n-2},w_{n-1})$

- History of size 1

  $P(w_1^n) \approx P(w_1)P(w_2|w_1)P(w_3|w_2)\ldots P(w_n|w_{n-1})$

- History of size 0

  $P(w_1^n) \approx P(w_1)P(w_2)P(w_3)\ldots P(w_n)$

Each of these are different models of language : tri-gram, bi-gram, uni-gram (bag-of-words)
Markov Models : $2^{nd}Order$, $1^{st}Order$, $0^{th}Order$

Independence Assumptions

- Uni-gram model
  - ∗ Each word is independent of the others in the sentence

    $P($all of a sudden I notice three guys standing on the sidewalk$) =$
    $P($sudden a three on sidewalk the standing $\ldots)$
- Bi-gram model
  - ∗ Each word is independent of all but the previous word.
- Tri-gram model
  - ∗ Each word is independent of all but the previous two words.

Put another way

- Put another way, trigram model assumes these are all equal:

  $P(\text{sidewalk}|\text{all of a sudden I notice three guys standing on the})$
  $P(\text{sidewalk}|\text{I walked on the})$
  $P(\text{sidewalk}|\text{See the small dog on the})$
  $P(\text{sidewalk}|\text{They carried out repairs on the})$

  because all are estimated as $P(\text{sidewalk}|\text{on the})$

  Not always a good assumption! But it does reduce the sparse data problem.

Estimating N-gram probabilities from a corpus

Maximum-Likelihood Estimation (MLE)

Unigram : $P(w_i) = \frac{count(w_i)}{N}$, where N is the number of words in the corpus.

Bigram : $P(w_i|w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$

bi-gram example: $P(\text{"roses"}|\text{"pink"}) = \frac{count(\text{"pink", "roses"})}{count(\text{"pink"})}$

Trigram : $P(w_i|w_{i-2}, w_{i-1}) = \frac{count(w_{i-2}, w_{i-1}, w_i)}{count(w_{i-2}, w_{i-1})}$

“Out of k occurrences of a context, how often is it followed by the given word”

Language Models

- Can compute the probability of a sequence of words

$$P(w_1, w_2, w_3 \dots w_n)$$

- Can compute the probability of an upcoming word, given previous context

$$P(w_i \mid w_1 \dots w_{i-1})$$

- Used in Surprisal based theories of language processing (Hale, 2001, Levy 2008)
  - $*$ processing difficulty of a word w in context c is proportional to its information-theoretic surprisal, defined as $-log\ p(w|c)$
- Used in neural Word Embeddings

Language Models

- Flexible framework: you do not have to always think of n-grams in terms of words
- Can also do n-grams of characters, phonemes (sound units), POS-tags, etc.