ML for Human Vision and Language

MSc Artificial Intelligence

Tejaswini Deoskar
t.deoskar@uu.nl

UiL-OTS, Utrecht University

Block 1, 2021-22

some slides adapted from S. Goldwater, A. Louis, A. Herbelot, and Jurafsky & Martin

MLHVL Part II

Machine Learning for Human Language

What does it mean to process language?

I've had a wonderful weekend! I always wanted to replace my old melodica.
On Saturday, I finally went to that fancy music store on Voorstraat.
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 + meaning of phrases, and sentences
- Resolve co-referring expressions
  - ∗ "it" refers to the melodica
- Relationships *between* sentences
  - ∗ I went *because* I wanted to buy a melodica
- Implicit information
  - ∗ I bought the melodica at that store on Voorstraat (in Utrecht)
  - ∗ The new melodica is the one I practised songs on...

Levels of Language Processing: Words

knowledge of words in the language and their meaning

This is a simple sentence WORDS

Morphology

knowledge of sub-word units of meaning

This	is	a	simple	sentence	**WORDS**
	be 3sg present				**MORPHOLOGY**

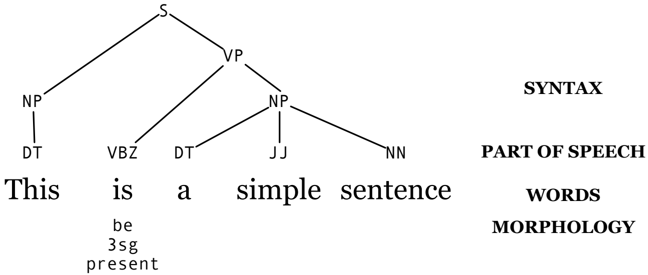
Parts of Speech

knowledge of 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**

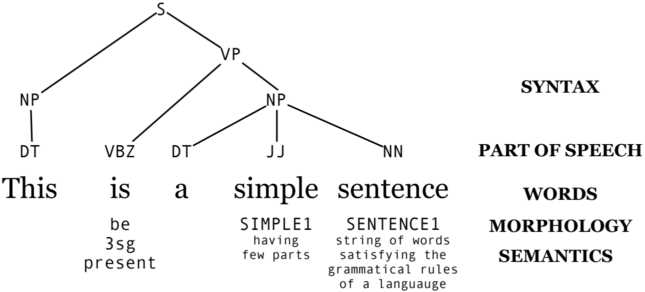
Grammar

structural relationships between words and phrases.

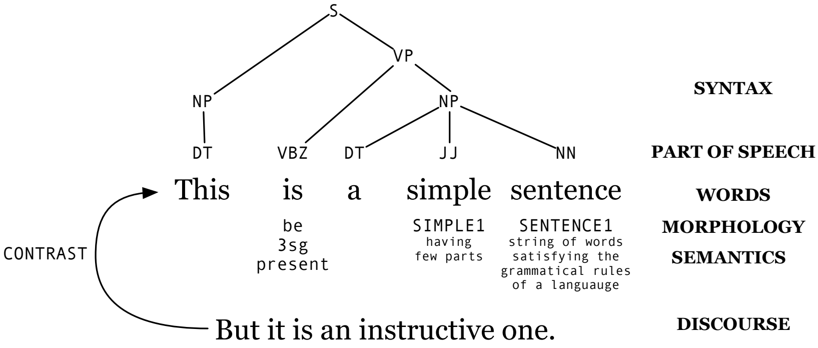


Meaning

Lexical (word) Semantics and Compositional Semantics



Discourse


S
VP
NP
NP
DT
VBZ
DT
JJ
NN
This is a simple sentence
be
3sg
present
SIMPLE1
having
few parts
SENTENCE1
string of words
satisfying the
grammatical rules
of a languauge
CONTRAST
But it is an instructive one.
SYNTAX
PART OF SPEECH
WORDS
MORPHOLOGY
SEMANTICS
DISCOURSE

Goals of language processing

- Scientific Goal: Build models of the human use of language
  - ∗ Humans are far 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.

Human vision versus language processing

- Vision: mapping between function and anatomy (e.g. cortical regions) is better defined
- Language: Only a very broad mapping possible so far
  - ∗ Indirect (non-invasive) methods like reading-time studies, eye-tracking studies
  - ∗ machine learning/ computational modelling of brain imaging data (FMRI, EEG)

Language processing and Deep Learning with ANNs

Chris Manning, *Stanford University*:

`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..... 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.`

Why is ML for Language hard?

Language is **ambiguous** at many levels.

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

(Number of ambiguities grow with sentence length, often in an exponential manner!)

Humans can disambiguate very well

.... but machines not so well!

A student in 2020 aced every course in AI.

The university sent a corona-test to every student. / The company sent a doctor to test every employee.

Why is ML for Language hard?

Did Google buy YouTube?

1. Google purchased YouTube
2. Google's acquisition of YouTube
3. Google acquired every company
4. YouTube is sold to Google
5. Google took over YouTube

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

Why is ML for Language hard?

Context dependence, world knowledge, and unknown representations

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

Topics : MLHVL part II

- **Distributed word meaning representations** (using non-neural + neural approaches) (lecture 5 + 6)
- Sequence modelling (RNN family) (lecture 7)
- Modelling of morphological learning in infants (using encoder-decoder NNs) (lecture 7)
- Combining image and language (lecture 8)

Organisation: Book chapter readings

Jurafsky and Martin (online edition)

- Vector Semantics (Chapter 6): All except 6.5 and 6.7.
- Background : Neural Networks (Chapter 7) up to 7.5
- Background : RNNs (Chapter 9) : up to 9.3 for sure (Rest if you have more experience in NNs/ have taken NLP course etc.)
- Encoder-decoder networks (Chapter 10) : all

Organisation: Lab assignments

- Lab assignments are in Pytorch (in the form of Jupyter Notebooks) : Group work
  - ∗ Pytorch Tutorial *(ungraded)*
  - ∗ 2 (A) : Learning Word embeddings (based on GloVe, Pennington et al., 2014.) *(graded, 10%)*
  - ∗ 2 (B) : RNN-based sequence model *(graded, 10%)*
- Individual assignment to be submitted at the end *(graded, 10%)*
- Check deadlines for everything on Blackboard.

Organisation: Research paper readings

- For Lab 2 (A)
  - ∗ Pennington J, Socher R, Manning C (2014) *GloVe: Global Vectors for Word Representation.*
- Brain imaging and word embeddings:
  - ∗ Mitchell TM et al (2008) *Predicting Human Brain Activity Associated with the Meanings of Nouns.* Science, 320: 1191-5.
  - ∗ Abnar et al. 2018 *Experiential, Distributional and Dependency-based Word Embeddings have Complementary Roles in Decoding Brain Activity*.
  - ∗ (Optional) *Artificial Neural Networks Accurately Predict Language Processing in the Brain*, Schrimpf et al. 2020.
- Modelling infant learning of morphology:
  - ∗ Maria Corkery, Yevgen Matusevych, and Sharon Goldwater, 2019. *Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection*

Lexical Semantics / Distributional Semantics

- Lexical semantics : how the meanings of individual words can be represented/learnt
- Distributional Semantics: a linguistic theory for representing word meaning based on contexts of use
  - ∗ DS word representations are a very flexible tool to model conceptual aspects of the lexicon
  - ∗ A variety of techniques used to produce/learn vectors
  - ∗ These vectors are used in countless applications
  - ∗ Good for open-class words : nouns, verbs, adjectives, adverbs.
  - ∗ Not so great for "Logical words" : *not, many, every, a, few ...*

A brief history of distributional approaches: From Wittgenstein to neural nets

Ludwig Wittgenstein: ‘Meaning is use’ : ’Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache? (Wittgenstein, 1953. 43)

J. R. Firth: ’You shall know a word by the company it keeps’ (Firth, J. R. 1957:11)

Karen Sparck-Jones: Early experiments on distributional semantics from thesauri for information retrieval, 1963, 1967.

The 2000s

- Larger corpora and better computational power start to give human-like representations of word meaning
- Psycholinguistic experiments show that distributional semantics can model linguistic phenomenon like similarly, priming, infant vocabulary acquisition.

The 2010s

- Neural models of distributional semantics slowly overtake non-neural models
- Distributional semantics enters neurolinguistics and is shown to be a good correlate of brain activation related to stored conceptual knowledge
- More and more advanced models: *contextualised word representations*
- We still can't do negation .... or composition!

Meaning from context(s)

What's the meaning of *'bardiwac'*?

- He handed her her glass of bardiwac.
- Beef dishes are made to complement the bardiwacs.
- Nigel staggered to his feet, face flushed from too much bardiwac.
- Malbec, one of the lesser-known bardiwac grapes, responds well to Australia's sunshine.
- I dined on bread and cheese and this excellent bardiwac.
- The drinks were delicious: blood-red bardiwac as well as light, sweet Rhenish.

⇒ *'bardiwac'* is a heavy red alcoholic beverage made from grapes

We are able to guess the meaning of *'bardiwac'* because the contexts in which it occurs (its distribution) are similar to the contexts where other known words appear:

similar distributions ⇔ *similar meaning*

Distributional Hypothesis

“The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear.” [Z. Harris (1954) *Distributional Structure*]

You shall know a word by the company it keeps (J. R. Firth, 1957)

Distributional Semantics

- Goal : find a representation that succinctly describes the meaning of a word (or phrase, sentence, document)
- Or at least : find a representation so as to determine if two words or texts have similar meanings

- Why is this a good thing?
  - ∗ Retrieve documents relevant to a query
  - ∗ Use representations as features for parsing, question answering, machine translation
  - ∗ Theory neutral, few assumptions re word meaning

Other approaches to lexical meaning

Formal semantics:

- meaning postulates, e.g.

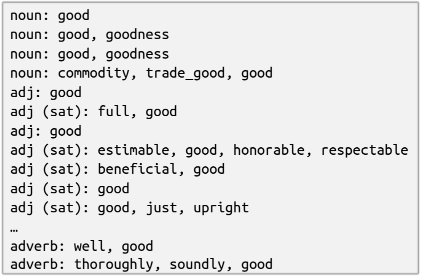
$$\forall x[bachelor'(x) \rightarrow male'(x) \wedge unmarried(x)]$$

- Defining concepts through enumeration of all of their features is highly problematic.
- Formal semantics useful for negation, quantification (every, some, all), etc.
- On-going research on combining distributional with formal/compositional semantics

Thesaurus based meaning

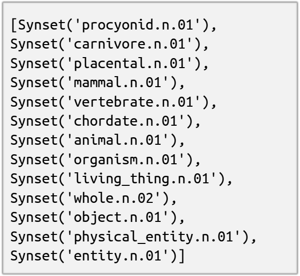
WordNet: a thesaurus containing lists of **synonym sets** and **hypernyms** (“X is a Y” relations)

synonym set containing good



```
noun: good
noun: good, goodness
noun: good, goodness
noun: commodity, trade_good, good
adj: good
adj (sat): full, good
adj: good
adj (sat): estimable, good, honorable, respectable
adj (sat): beneficial, good
adj (sat): good
adj (sat): good, just, upright
…
adverb: well, good
adverb: thoroughly, soundly, good
```

hypernyms of panda



```
[Synset('procyonid.n.01'),
Synset('carnivore.n.01'),
Synset('placental.n.01'),
Synset('mammal.n.01'),
Synset('vertebrate.n.01'),
Synset('chordate.n.01'),
Synset('animal.n.01'),
Synset('organism.n.01'),
Synset('living_thing.n.01'),
Synset('whole.n.02'),
Synset('object.n.01'),
Synset('physical_entity.n.01'),
Synset('entity.n.01')]
```

Problems with resources like WordNet

- Great as a resource but not fine-grained enough
  - ∗ e.g. proficient is listed as a synomym of good, but this is correct only in certain contexts
- Missing new meanings
  - ∗ e.g. wicked, badass
- Subjective ; Require human labour
- Cannot compute word similarity effectively.
- Still widely used; **retrofitting** word-vectors using wordnet helps

Problem with representing words as discrete symbols

- Traditionally, we regard words as discrete symbols:
  hotel, motel

- Words can be represented by one hot vectors

  motel = [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]
  hotel = [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0]

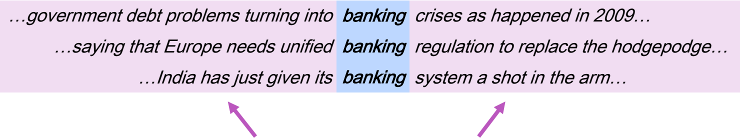
  [Vector dimension = number of words in vocabulary (e.g. 50,000)]

- No natural notion of similarity for one-hot vectors! These two vectors are orthogonal.

- **Instead**: learn to encode similarity in the vectors themselves!

Representing words by their contexts

- When a word w appears in a text, its context is the set of words that appear in its vicinity



…government debt problems turning into ***banking*** *crises as happened in 2009…*

…saying that Europe needs unified ***banking*** *regulation to replace the hodgepodge…*

…India has just given its ***banking*** *system a shot in the arm…*

These context words will represent ***banking***

- Use the many contexts of w to build up a vector representation of w

	regulation	system	crises	debt	pet	bone	fur	fetch
banking =	1	1	1	1	0	0	0	0
⋮								
dog =	0	0	0	0	1	1	1	1

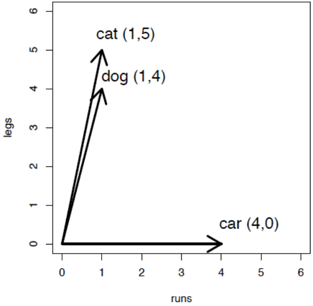
- context words are the *dimensions* of the vector (here boolean valued)

The semantic space

- Vectors exist in a vector space - the semantic space.

A toy space with 3 word vectors in 2 dimensions:

	run	legs
dog	1	4
cat	1	5
car	4	0



- The semantic space has dimensions that correspond to possible contexts (also called features of the semantic space).

- A distribution is a point in this space (the vector is defined wrt the origin of that space)

Learning the semantic space

Distributional semantics: *family of techniques* for representing word meaning based on (linguistic) contexts of use.
Two main types of techniques for learning:

1. Count-based models:
   - count the frequency of co-occurrence of a target and a context
   - words are represented as vectors (sparse)
2. Predict-based models:
   - try to predict plausible contexts for a word
   - learn word representations in the process (dense)
   - generally implemented as a neural network
3. GloVe: a count-based method with benefits!

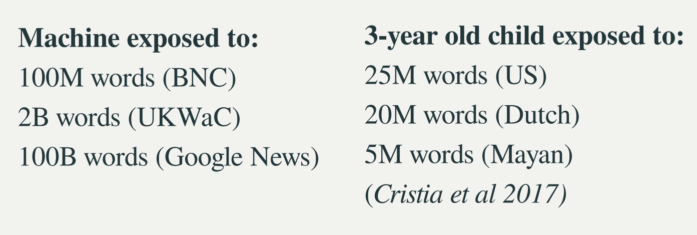
Learning the semantic space

- Modelling choices
- How to measure similarity
- Evaluation of the learnt semantic space/models

Modelling choices: how much data?

- as much data as possible?
  - ∗ British National Corpus (BNC) : 100M words
  - ∗ BERT, 2019: BookCorpus (800M words) + Wikipedia (2.5G words)
- in general, more data is preferable, but
  - ∗ more data is not realistic from a psycholinguistic point
  - ∗ Humans don't get exposed to so much data

Vectors vs human meaning



Machine exposed to:	**3-year old child exposed to:**
100M words (BNC)	25M words (US)
2B words (UKWaC)	20M words (Dutch)
100B words (Google News)	5M words (Mayan)
	(*Cristia et al 2017)*

slide from A. Herbelot

Modelling choices: how to define context

- Word windows (unfiltered) : n words on either side of the lexical item.

  Example: $n = 2$ (5 word window):

  ...[needs unified **banking** regulations to]...

- Word windows (filtered): n words on either side removing some words, e.g. function words, other very frequent /uninformative words (**stopwords**)

  Example: $n = 2$ (5 word window):

  ...[needs unified **banking** regulations (to) replace]...

- Lexeme window: As above, but use *stems* (base forms) of words.

  ...[need unify **bank** regulation (to) replace]...

Modelling choices: how to define context

- Part of speech tags : Include the parts of speech of words
- Dependencies: Context for a word might be the dependency structure it belongs to (Pado and Lapata, 2007)

The prime ***minister*** *acknowledged the question.*

minister [prime_a 1, acknowledge_v 1]

minister [prime_a_mod 1, acknowledge_v_subj 1]

minister [prime_a 1, acknowledge_v+question_n 1]

Modelling choices: context size

- Choice of context size might depend on the kind of vectors you want
  - ∗ For *semantic* similarity, use a large window (50 words)
  - ∗ For *syntactic* similarity, use a small window (1-3 words)
  - ∗ For *document* similarity, use entire document
- Perceptual information: images, sounds, etc. to complement the conceptual data given by language alone

Modelling choices : Context weighting

- Binary model : if context c co-occurs with word w, value of vector $\vec{w}$ for dimension c is 1, 0 otherwise.
- Basic frequency model: value of vector $\vec{w}$ for dimension c is the number of times c co-occurs with w
- Is Basic frequency good enough?
  - $*$ raw corpus data and frequencies do *not* model human semantic judgements well
  - $*$ raw statistical information from the corpus is modified (using context weighting or dimensionality reduction or a mix)

Modelling choices : Context weighting

- frequent words are expected to have high counts in the context vector
  - ∗ regardless of whether they occur more often with one word than with others
- We want to know which words occur **unusually** often in the context of a target word w: more than we'd expect by chance.
  - ∗ Collocations: words or terms that co-occur more often than would be expected by chance

Pointwise Mutual information (PMI)

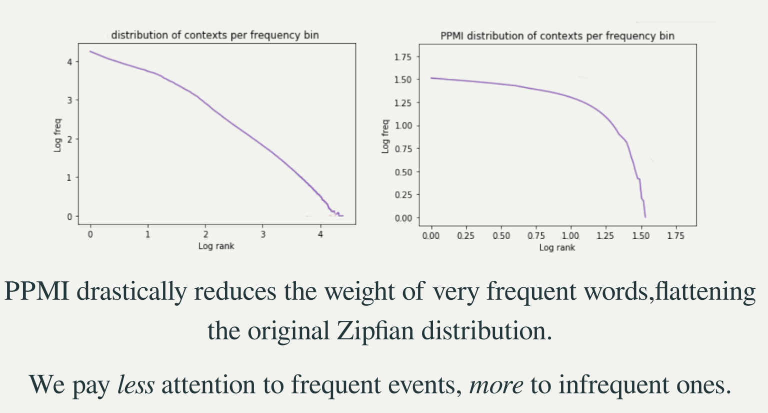
- One way: use pointwise mutual information:

$$PMI(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)}$$

  - ∗ Numerator: Actual prob. of seeing words x and y together
  - ∗ Denominator: Predicted prob. of same, if x and y are independent.

- Positive PMI : x and y are more likely to occur together than if independent. e.g. pepperoni and pizza
- Negative PMI : x and y are less likely to occur together than if independent. e.g. AI and pizza
- Good collocation pairs have high PMI
- A derivative of PMI (Positive PMI) used in practise

De-Zipfianising with PPMI



PPMI drastically reduces the weight of very frequent words,flattening the original Zipfian distribution.

We pay *less* attention to frequent events, *more* to infrequent ones.

- Contrary to the ‘pure’ view of distributional semantics (*You shall know a word by the company it keeps*, raw corpus data does not model human semantic judgements very well!