Natural Language Models and Interfaces

BSc Artificial Intelligence

Lecturer: Wilker Aziz
Institute for Logic, Language, and Computation

2018, week 1, lecture a

NLMI

Course organisation

Why NLP?

Why is NLP hard?

An overview of problems

An overview of the statistical method

Language data: first contact

Course

Topic: Statistical Natural Language Processing

Team

- Instructor: Wilker Aziz
- Assistants: Miguel Rios, Urja, Evelyne, Nora, Caitlin, Tessa

Attendance

- lectures: not monitored, but encouraged
- labs: highly encouraged!

Course information

Blackboard

- course manual
- weekly materials: readings, slides
- assignments: exercises, projects

Textbook
Jurafsky & Martin, *Speech and Language Processing* (2nd edition)

Additional material may be announced in class

Assessment

- One exam (individual work): 40%
- Weekly assignments (in pairs)
  - practical programming exercises and mini-projects: 40%
  - theoretical (non-programming) exercises: 20%

NLMI

What about processing language?

It's everywhere!

- We talk about things
  I love Paris! All those bridges, the cathedral, the Louvre, oh and of course, the tower!

What about processing language?

It's everywhere!

- We talk about things
  I love Paris! All those bridges, the cathedral, the Louvre, oh and of course, the tower!
- We give instructions
  From Dam square you head north on Damrak till you see it, really, you can't miss it.

What about processing language?

It's everywhere!

- We talk about things
  I love Paris! All those bridges, the cathedral, the Louvre, oh and of course, the tower!
- We give instructions
  From Dam square you head north on Damrak till you see it, really, you can't miss it.
- We entertain ourselves

Eleanor Ribgy

... picks up the rice
In the church where a wedding has been
Lives in a dream
Waits at the window, wearing the face
That she keeps in a jar by the door
Who is it for

People infer stuff from text and speech

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.

Adapted from A. Louis, S. Goldwater, I. Titov, K. Sima'an, T. Deoskar

People infer stuff from text and speech

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.

- meaning of words, phrases, and sentences
  A melodica is a musical instrument, Haarlem is a place

Adapted from A. Louis, S. Goldwater, I. Titov, K. Sima'an, T. Deoskar

People infer stuff from text and speech

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.

- meaning of words, phrases, and sentences
  A melodica is a musical instrument, Haarlem is a place
- relationships between sentences
  I went *because* I wanted to buy a melodica

Adapted from A. Louis, S. Goldwater, I. Titov, K. Sima'an, T. Deoskar

People infer stuff from text and speech

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.

- meaning of words, phrases, and sentences
  A melodica is a musical instrument, Haarlem is a place
- relationships between sentences
  I went *because* I wanted to buy a melodica
- implications
  The melodica was bought at that store in Haarlem

Adapted from A. Louis, S. Goldwater, I. Titov, K. Sima'an, T. Deoskar

People infer stuff from text and speech

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.

- meaning of words, phrases, and sentences
  A melodica is a musical instrument, Haarlem is a place
- relationships between sentences
  I went *because* I wanted to buy a melodica
- implications
  The melodica was bought at that store in Haarlem
- impressions about speaker/writer style
  The writing is boring or funny or engaging

Adapted from A. Louis, S. Goldwater, I. Titov, K. Sima'an, T. Deoskar

All of this understanding plays a role when we

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

NLP then is about enabling computers to do some of these tasks

- How to study/analyse language in computational terms?
- How to build applications that will do these tasks automatically?

Adapted from A. Louis, S. Goldwater, I. Titov, K. Sima'an, T. Deoskar

Goals of NLP

Scientific

- Build models of the human use of language

Engineering

- Build models that serve in technological applications
  - machine translation
  - speech systems
  - information extraction, etc.

Goals of NLP

Scientific

- Build models of the human use of language

Engineering

- Build models that serve in technological applications
  - machine translation
  - speech systems
  - information extraction, etc.

In this course we

- draw insights from scientific knowledge
- but mostly focus on engineering aspects
- and rely on language data in the form of digital text

NLP Applications

- Information retrieval: Google
- Summarisation: Google News
- Speech recognition: Siri, Alexa, Google Home
- Dialogue systems: Amazon chatbot
- Machine translation: Google translate
- Image captioning: Microsoft, Facebook
- Recommendation systems: Amazon reviews
- Social network analysis: Facebook, Twitter

NLMI

Course organisation

Why NLP?

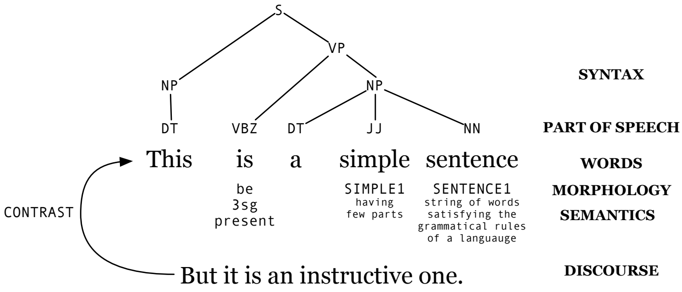
Why is NLP hard?

An overview of problems

An overview of the statistical method

Language data: first contact

Basic levels of structure



Slide from S. Goldwater

Why is NLP hard?

Ambiguity at many levels

- Word senses: bank (finance or river?)
- Part of speech: chair (noun or verb?)
- Syntactic structure: I saw a man with a telescope
- Quantifier scope: Every child loves some movie
- Multiple: I saw her duck

and ambiguity typically grows with sentence length

Adapted from T. Deoskar

Why is NLP hard?

Ambiguity at many levels

- Word senses: bank (finance or river?)
- Part of speech: chair (noun or verb?)
- Syntactic structure: I saw a man with a telescope
- Quantifier scope: Every child loves some movie
- Multiple: I saw her duck

and ambiguity typically grows with sentence length

Examples from newspaper headlines

Iraqi head seeks arms
Stolen painting found by tree
Teacher strikes idle kids

Adapted from T. Deoskar

Why is NLP hard?

Variability (paraphrasing)

- *Emma burst into tears and he tried to comfort her, saying things to make her smile.*
- *Emma cried, and he tried to console her, adorning his words with puns.*

Example from **?**

Why is NLP hard?

Different genres

- Suppose we train a part of speech tagger on the Wall Street Journal

  Mr./NNP Vinken/NNP is/VBZ chairman/NN of/IN Elsevier/NNP N.V./NNP ,/, the/DT Dutch/NNP publishing/VBG group/NN ./.

- What will happen if we try to use this tagger for social media??

  ikr smh he asked fir yo last name

Twitter example due to Noah Smith

Why is NLP hard?

Languages are different

- Chinese sentences do not have delimiters between words

(a) Raw data:

他还提出一系列具体措施和政策要点。

(b) Segmented:

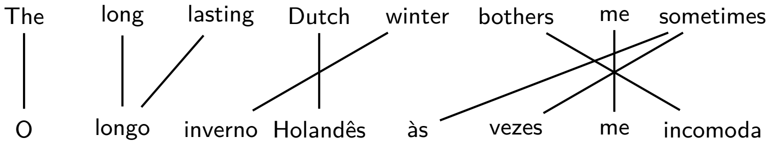
他	还	提出	一	系列	具体	措施	和	政策	要点	。
He	also	propose	one	series	concrete	measure	and	policy	essential	.

(He also proposed a series of concrete measures and essentials on policy.)

Example from **?**

Why is NLP hard?

Languages have different word orders



Myself (2018)

Why is NLP hard?

Context dependence

- correct interpretation typically requires context and often requires world knowledge
  Paris is so beautiful, the city or the celebrity?

Why is NLP hard?

Context dependence

- correct interpretation typically requires context
  and often requires world knowledge
  Paris is so beautiful, the city or the celebrity?
  especially when seen from the tower

Why is NLP hard?

Context dependence

- correct interpretation typically requires context
  and often requires world knowledge
  Paris is so beautiful, the city or the celebrity?
  especially when seen from the tower

Unknown representation

- we don't know how humans represent knowledge

NLMI

Course organisation

Why NLP?

Why is NLP hard?

An overview of problems

An overview of the statistical method

Language data: first contact

Sequential prediction

What is the next word? ▸ quiz

- I slept on my ...
- Where is the ...
- Natural language processing is ...

Sequential prediction

What is the next word? quiz

- I slept on my . . .
- Where is the . . .
- Natural language processing is . . .

not every word is equally likely to continue a certain prefix

- we typically make meaningful and grammatical sentences

Sequence segmentation

Some languages are based on *continuous scripts* Wiki

- for example Chinese and Thai

In English, words are generally clearly delimited

- but we still care about tokenisation
  - input: I'm not missing it, neither should ya!
  - output: I 'm not missing it , neither should ya !

quiz

Sequence segmentation

Some languages are based on *continuous scripts* Wiki

- for example Chinese and Thai

In English, words are generally clearly delimited

- but we still care about tokenisation
  - input: I'm not missing it, neither should ya!
  - output: I 'm not missing it , neither should ya !

quiz

It is not necessarily clear what it means to find a segmentation

- we are either looking for meaning carrying parts
- or trying to minimise the cost of representation

Sequence labelling

We are often interested in analysing sentences

- we can classify words with respect to parts of speech
  apple is a noun
- and context usually plays a role
  I chair$_{\text{verb}}$ debates all the time, and usually I do not have a chair$_{\text{noun}}$ to sit on
- some words may refer to an entity
  Leibniz Wiki *was a German mathematician*

It's similar to sequence prediction, but with additional context

quiz

- it may require far more knowledge of the world

Morphological disambiguation

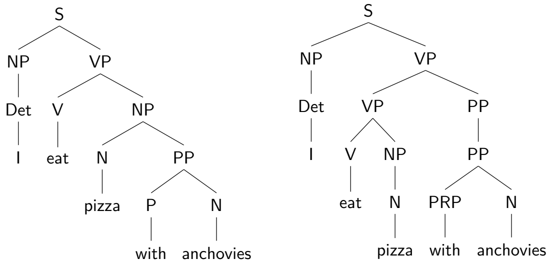
Words have meaning carrying and functional parts

- English -ly usually *derives* an adverb from an adjective
- less often English can use *agglutination* or *compounding* to make new words
  wrongdoing is wrong + doing
- there are ambiguities
  - s marks plural in *cat*s, third person in *it mark*s, nothing in *new*s
  - with a verb un means "reversal", e.g. un*tie*
    with an adjective un means "not", e.g. un*wise*
- other languages are far more complex ▸ Wiki

Syntactic parsing

We can take the idea of sequence labelling and push it a bit farther

- label every "coherent" substring in a sentence
  a constituent
- and we can do so recursively



which one has a funny interpretation?

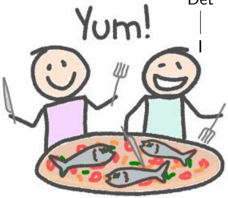
Stanford parser demo ▸ Try it out!

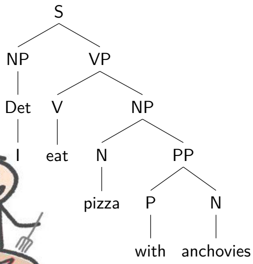
Drawings from James Constable — CC 2.5

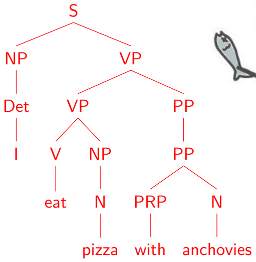
Syntactic parsing

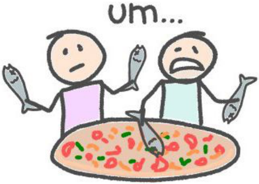
We can take the idea of sequence labelling and push it a bit farther

- label every "coherent" substring in a sentence
  a constituent
- and we can do so recursively









which one has a funny interpretation?

nesting tells us about syntactic dependencies

Stanford parser demo ▸ Try it out!

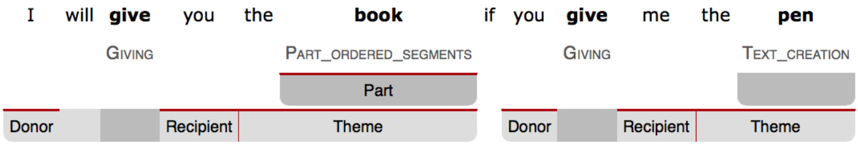
Drawings from James Constable — CC 2.5

Semantic parsing

We may be interested in the semantic role of constituents
with respect to a predicate ▸ Wiki
rather than their syntactic function

Answer questions such as

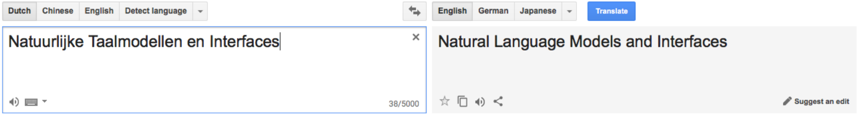
- *who* did *what* to *whom*, *when* and *why*?



ARK's Syntactic & Semantic Parsing Demo ▸ Try it out!

Text-to-text transformation

We can combine sequential prediction with sequence labelling and a few more things to translate ▸ seq2seq



or summarise

Google seq2seq

Much more

- coreference resolution
- discourse analysis
- question answering
- paraphrasing
- translation equivalence
- word alignment

NLMI

Course organisation

Why NLP?

Why is NLP hard?

An overview of problems

An overview of the statistical method

Language data: first contact

But how can we do that?

Statistical approach

- or the "probabilistic pipeline"

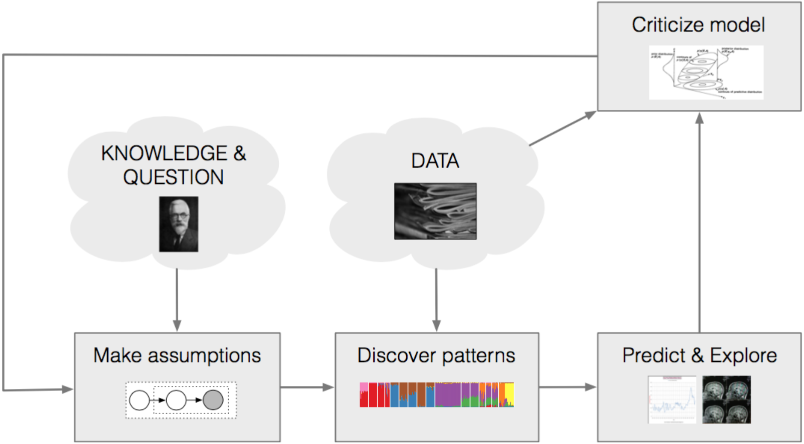


Image by David Blei

Pipeline

We have knowledge about the world and we have questions we want to answer

- so we can design a model: encodes our knowledge and assumptions

Pipeline

We have knowledge about the world and we have questions we want to answer

- so we can design a model: encodes our knowledge and assumptions

We have data that by assumptions is somewhat complies with our assumptions

- so we can use statistics to discover patterns in data

Pipeline

We have knowledge about the world and we have questions we want to answer

- so we can design a model: encodes our knowledge and assumptions

We have data that by assumptions is somewhat complies with our assumptions

- so we can use statistics to discover patterns in data

We typically want to predict things or explore things

- again statistics can help us make decisions
- predict future outcomes
- organise unstructured data is some structured way

What do people talk about in the Wall Street Journal?

1	2	3	4	5
Game	Life	Film	Book	Wine
Season	Know	Movie	Life	Street
Team	School	Show	Books	Hotel
Coach	Street	Life	Novel	House
Play	Man	Television	Story	Room
Points	Family	Films	Man	Night
Games	Says	Director	Author	Place
Giants	House	Man	House	Restaurant
Second	Children	Story	War	Park
Players	Night	Says	Children	Garden

6	7	8	9	10
Bush	Building	Won	Yankees	Government
Campaign	Street	Team	Game	War
Clinton	Square	Second	Mets	Military
Republican	Housing	Race	Season	Officials
House	House	Round	Run	Iraq
Party	Buildings	Cup	League	Forces
Democratic	Development	Open	Baseball	Iraqi
Political	Space	Game	Team	Army
Democrats	Percent	Play	Games	Troops
Senator	Real	Win	Hit	Soldiers

11	12	13	14	15
Children	Stock	Church	Art	Police
School	Percent	War	Museum	Yesterday
Women	Companies	Women	Show	Man
Family	Fund	Life	Gallery	Officer
Parents	Market	Black	Works	Officers
Child	Bank	Political	Artists	Case
Life	Investors	Catholic	Street	Found
Says	Funds	Government	Artist	Charged
Help	Financial	Jewish	Paintings	Street
Mother	Business	Pope	Exhibition	Shot

Topics found in 1.8M articles from the New York Times

[Hoffman+ JMLR 2013]

NLMI

Course organisation

Why NLP?

Why is NLP hard?

An overview of problems

An overview of the statistical method

Language data: first contact

Let's start with the frequency of words

There are always phenomena which are important but have rare evidence in data: Zipf's Law.

- To illustrate, let's look at the frequencies of different words in a large text corpus.
- Assume a "word" is a string of letters separated by spaces (a great oversimplification as we know by now)

Adapted from T. Deoskar

Word Counts

Most frequent words in the English Europarl corpus
out of 24 million 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

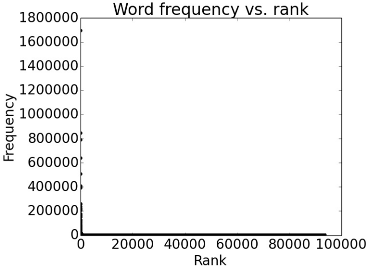
Out of 93638 distinct words (word types), 36231 occur only once!

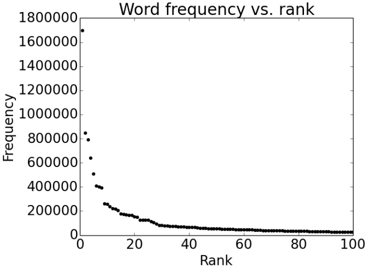
Examples:

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

Plotting word frequencies

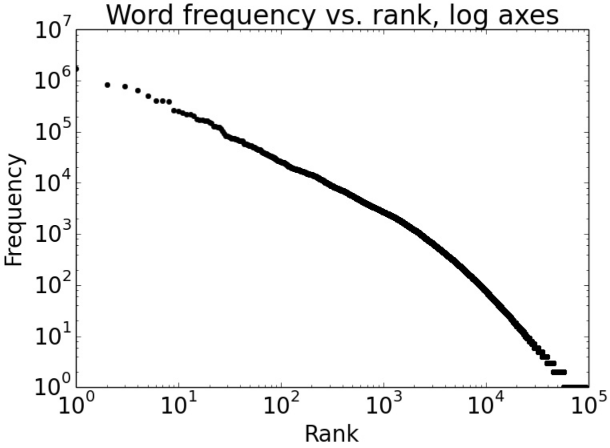
If we order words by frequency,
what is the frequency of nth ranked word?





Rescaling the axes

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



Zipf's law

Summarises the behaviour we just saw:

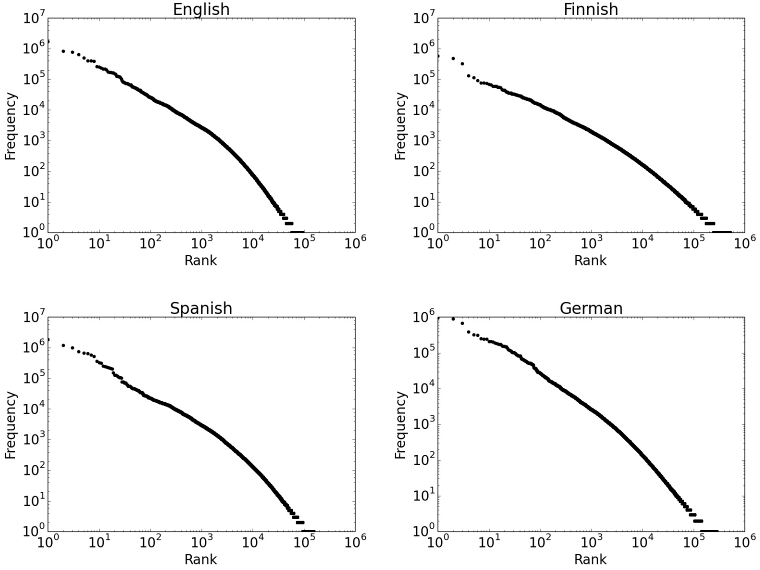
$$f \times r \approx k$$

- $f =$ 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

- 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 we need to find clever ways to estimate probabilities for things we have rarely or never seen.

Scope of the course

In this course you will learn about

- probabilistic modelling
- statistical inference and estimation
- how to represent language data
- discovering patterns in text collections

Topics

- Markov models: including language models and sequence prediction
- Mixture models: sequence labelling and PCFGs
- Models of distributional semantics: word representation
- Translation equivalence: learning dictionaries

Topics

- Markov models: including language models and sequence prediction
- Mixture models: sequence labelling and PCFGs
- Models of distributional semantics: word representation
- Translation equivalence: learning dictionaries

See you next time for

- a review of probabilities and parameter estimation

References I