Queen Mary
University of London

ECS763 Natural Language Processing

Unit 1: NLP Applications and properties of language

Lecturer: Julian Hough
School of Electronic Engineering and Computer Science

OUTLINE

1) What is NLP and where is it used?
2) A typical application: sentiment analysis
3) Properties of natural language and intro to mathematical methods
   3.1) The word level
   3.2) The sentence level
   3.3) Beyond the sentence level for discourse
   3.4) Intro to dialogue and its challenges

OUTLINE

1) What is NLP and where is it used?
2) A typical application: sentiment analysis
3) Properties of natural language and intro to mathematical methods
   3.1) The word level
   3.2) The sentence level
   3.3) Beyond the sentence level for discourse
   3.4) Intro to dialogue and its challenges

What is Natural Language Processing?

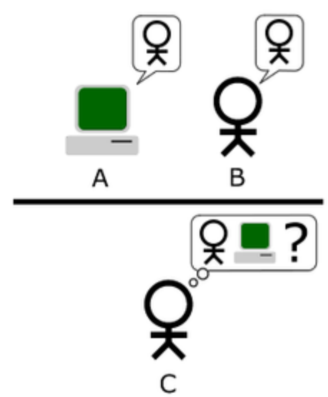
Natural Language Processing

(or Computational Linguistics)

is the automatic processing of human language data for different purposes.

What is Natural Language Processing?

- **BIG PICTURE 1:** We really want to build machines that **understand** human language in a human way, and **produce/generate** human language in a human way.
- Alan Turing (1950) originally posed the **Turing Test** as being key to solving artificial intelligence (AI).
- Could a machine ‘fool’ someone into thinking they’re talking to a human? That system will have solved AI**.**



What is Natural Language Processing?

- <u>**BIG PICTURE 2:**</u> We want **tools** that allow us to do tasks more effectively.
- This technology might assist you with **organising** huge amounts of text information, accessing parts of it, and extracting data from it.
- It can help you **create** your own text data: e.g. spelling and style correction.
- It can **help/assist** those who need it to **complete tasks:** text-to-speech from screens for the blind or drivers; speech-to-text for those with manual problems; saves labour in call centres etc.

What is Natural Language Processing?

- **Why** is it worth studying?
  - Huge number of applications to help humans do useful tasks.
  - Consequently has huge commercial and social value.
  - Theoretical interest as it shines a light on how human beings use language to communicate.

- As a **field** it's at the intersection of:
  - Computer Science
  - Data Science
  - AI / Machine Learning (More recently Deep Learning)
  - Linguistics / Cognitive Science

Why is NLP difficult and interesting? Because human language is…

- **Ambiguous (can mean several things at once)** (unlike programming languages)

- **Not always explicit and depends on context**. You leave out “code”- the listener/reader fills in the gaps!
  - **Context** includes real-world knowledge. Do words ‘mean’ anything without reference to real things/situations?

- **Rich** in its ability to express lots of things.

- **Creative and free**- you can always create a new word/ phrase!

Applications: main areas

- Machine Translation (since the 1950s)
- **Managing BIG data**:
  - Search (Google)
  - Analysing social media for advertising e.g. **sentiment analysis** for products.
  - Finance: buy/sell decisions based on social media texts. Health: Which hospitals are good?
- **Dialogue systems/Chatbots**:
  - Personal assistants (Amazon's Alexa, Apple's Siri).
  - Human-robot interaction with speech.
  - Automating customer service.

Applications (simple to complex)

- Keyword search
- Spell-checking/auto-complete
- Extracting information from websites: product, price, company names
- Summarization of texts
- Classification: sentiment classification (positive or negative), difficulty of reading level of text
- Machine translation
- Question Answering
- Conversation Analytics
- Dialogue Systems (spoken and typed interfaces)

Applications: Machine translation

- The earliest form of NLP. Started in the 1950s.
- Now widely used with large scale statistical methods.
- Now large scale MT (e.g. Google translate) is pretty impressive, with a huge number of language pairs.
- “The Google Translate app supports more than 100 languages and can translate via photo, via voice in "conversation mode”, and via real-time video in "augmented reality mode".”

Applications: Managing big (textual) data

- **CLASSIFY** text so as to identify relevant content / quickly assess this content
  - E.g., **Sentiment Analysis, Hate Speech Detection from Social Media posts, Spam Detection**

- **EXTRACT** structured information from unstructured textual data

- **SUMMARIZE** text- compressing the full text into smaller readable summaries

Classification

From: "" <takworlld@hotmail.com>

Subject: real estate is the only way... gem oalvgkay

Anyone can buy real estate with no money down

Stop paying rent TODAY !

There is no need to spend hundreds or even thousands for similar courses

I am 22 years old and I have already purchased 6 properties using the methods outlined in this truly INCREDIBLE ebook.

Change your life NOW !

===

Click Below to order:

http://www.wholesaledaily.com/sales/nmd.htm

===

Spam

or

not Spam?

Classification

`i love @justinbieber #sarcasm` → Positive or Negative?

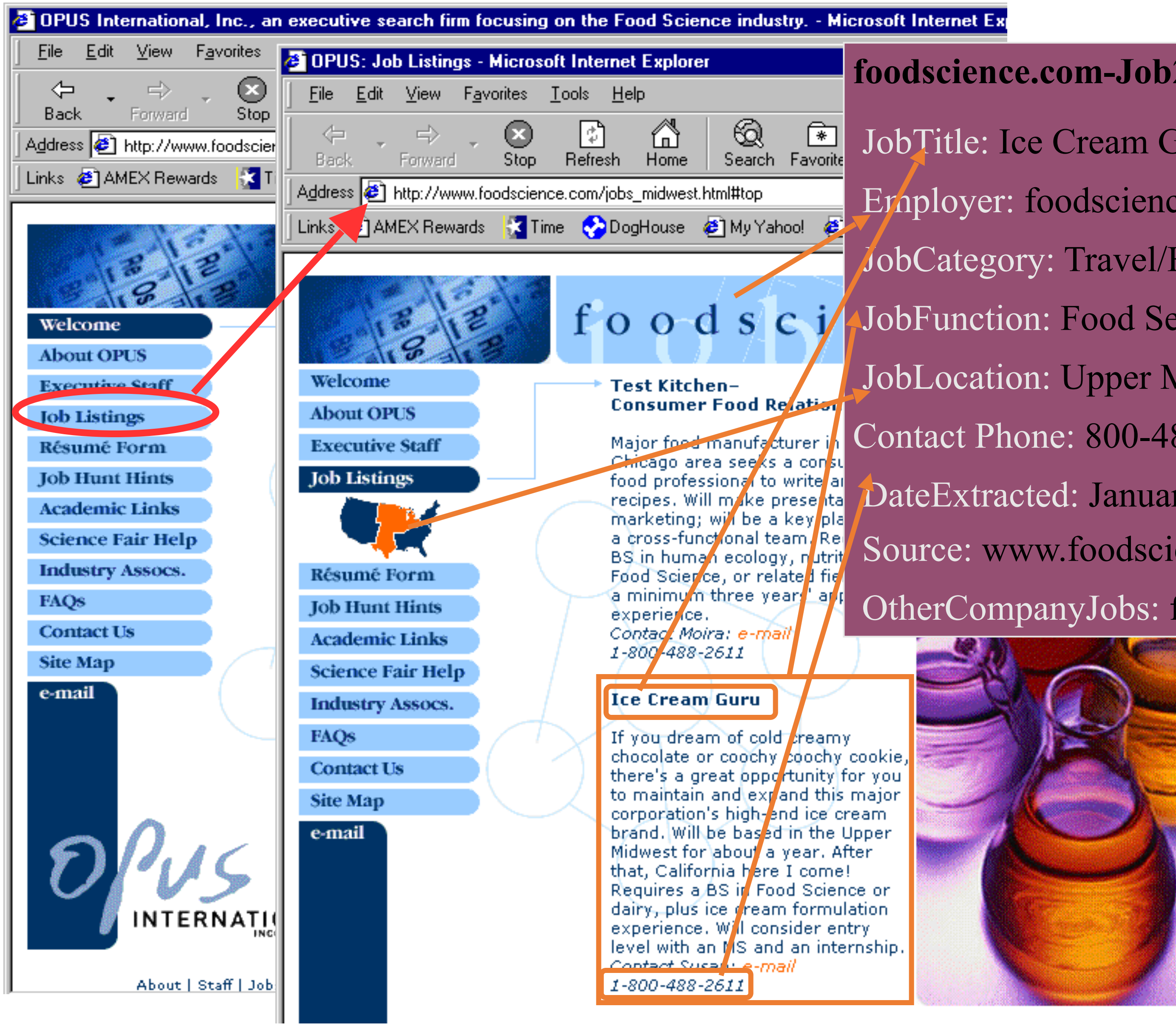
Classification

Hate

or

Non-hate?

Information Extraction



foodscience.com-Job2

JobTitle: Ice Cream Guru

Employer: foodscience.com

JobCategory: Travel/Hospitality

JobFunction: Food Services

JobLocation: Upper Midwest

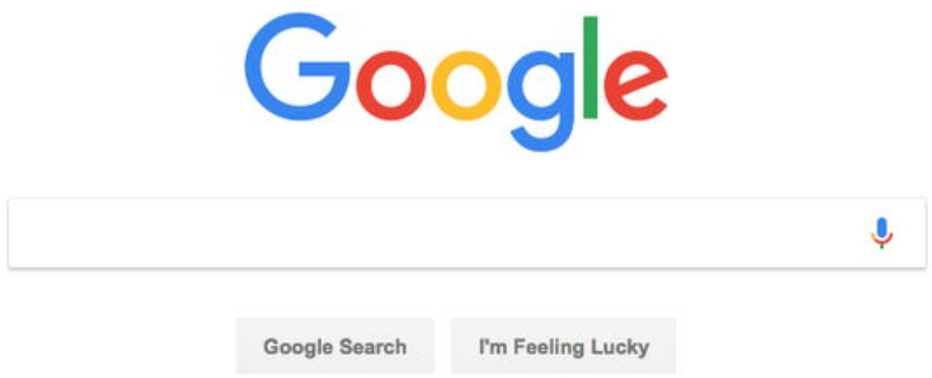
Contact Phone: 800-488-2611

DateExtracted: January 8, 2001

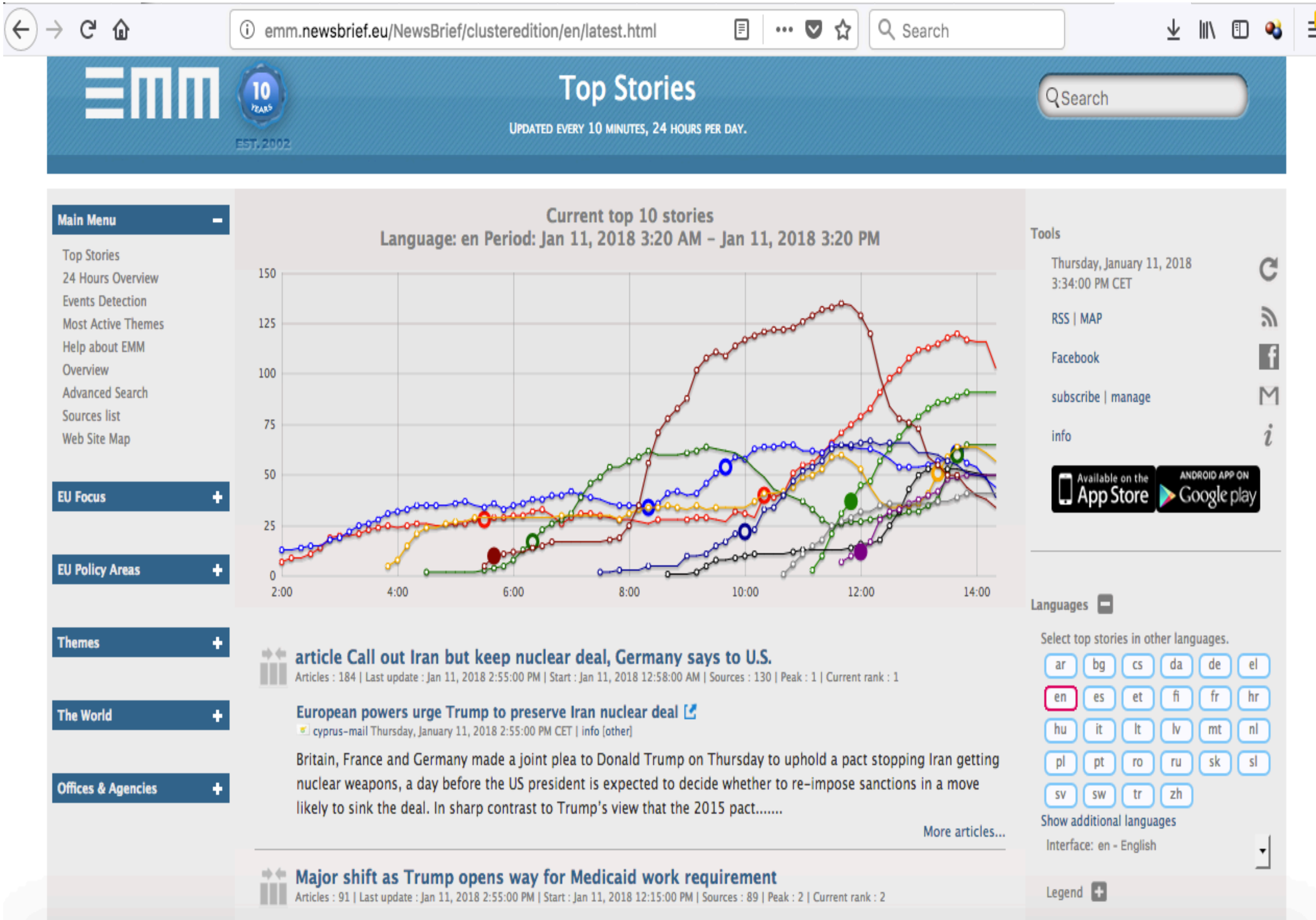
Source: www.foodscience.com/jobs_midwest.html

OtherCompanyJobs: foodscience.com-Job1

Information Extraction



Summarization



Summarization

- **Summarization** is the production of a summary either from a single source (single-document summarization) or from a collection of articles (multi-document summarization)

- Main approaches are:
  - **Extractive**: Select key sentences/phrases for summary.
  - **Abstractive**: Re-generate a summary based on the meaning of the text.

Applications: Dialogue systems

Applications: Dialogue systems

- The advent of mobile phones has been a blessing to NLP for commercial systems.

- Gave rise to Siri, then Google Assistant, Cortana. Question Answering and information retrieval through voice.

- Then finally it has adopted into people’s homes- Alexa, Google Home.

Applications: Dialogue systems

- Chatbots (text-based)
  - Personal assistants
  - Online helpline/FAQ answering
  - Helps reduce human labour
  - Google DialogFlow is an easy open source toolkit to build chatbots **(Unassessed exercise on this)**

- Spoken dialogue systems (speech-based)
  - Artificial call centre employees
  - In robots/cars
  - Can be artificial companions and again, helps reduce human labour

OUTLINE

1) What is NLP and where is it used?
2) A typical application: sentiment analysis
3) Properties of natural language and intro to mathematical methods
    3.1) The word level
    3.2) The sentence level
    3.3) Beyond the sentence level for discourse
    3.4) Intro to dialogue and its challenges

Sentiment Analysis

1. Id: Abc123 on 5-1-2008 “I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too.

2. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …”

Sentiment Analysis

POSITIVE about IPhone ☺️

1. Id: Abc123 on 5-1-2008 “I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too.

2. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …”

Sentiment Analysis

POSITIVE about IPhone ☺️

1. Id: Abc123 on 5-1-2008 “I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too.

2. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …”

POSITIVE about IPhone ☺️

NEGATIVE about Blackberry ☹️

Sentiment Analysis

- A typical NLP task- a type of **classification** task.
- You have a large amount of data available to you (a **corpus**). E.g. collection of tweets or comments.
- You need to build something to make the automatic decision about a text (tweet) to **classify** it as:
  - **Positive** ☺️ **vs Negative** ☹️

```
I'm really happy!
I'm having a terrible day
Oh man this is so great <3
I just can't believe it
```

- **<u>How could we go about this?</u>**

Sentiment Analysis First attempt: Dictionaries

- We could build dictionaries:
  - List of “positive” words
  - List of “negative” words

- **<u>What is wrong with this approach/what are its limitations?</u>**

- Problem with ambiguity- is this positive or negative?:

```
i love @justinbieber #sarcasm
```

- We might need a more data-driven approach…

Sentiment Analysis: Data-Driven Classification

- We could **learn** the dictionaries of ‘positive’ and ‘negative’ words from **data,** which would have:
  - A list of “positive” examples
  - A list of “negative” examples

Negative ☹️

```
I'm having a terrible day
I just can't believe it
I love @justinbieber #sarcasm
```

Positive ☺️

```
I'm really happy
Oh man this is just so great <3
I love @justinbieber
```

- **Train** a **classifier** to label a text as positive or negative based on observed words and combinations thereof in each example, then **test it** on unseen examples to see how good it is.
- We can use **probability, statistics** and **geometry.**

Sentiment Analysis: Data-Driven Classification - Preprocessing

- We're going to have to use the words from the texts to feed into our classifier- but what else is there?
- But how do actually we get to the words? i.e. what **pre-processing**?
- At least:
  - Sentence **segmentation**
    - (split? At what?)
  - Word **tokenisation**
    - (split? At what? Just standard words or something else?)
- And maybe:
  - **Normalisation**, **spelling correction**
    - (how?)
  - **Stop word** removal
    - (really?)

Tokenisation

- We need to define which tokens of each text to use.
- Issues in tokenisation:
  - ***Finland's capital*** →
    Finland? Finlands? Finland's?
  - ***Hewlett-Packard*** → ***Hewlett*** and ***Packard*** as two tokens?
    - ***state-of-the-art***: break up hyphenated sequence.
    - ***co-education***
    - ***lowercase***, ***lower-case***, ***lower case*** ?
    - It's effective to get the user to put in possible hyphens
  - ***San Francisco***: one token or two? How do you decide it is one token?

Normalisation

- Need to “normalise” terms in indexed text as well as query terms into the same form
  - We want to match ***U.S.A.*** and ***USA***
- We most commonly implicitly define equivalence classes of terms
  - e.g., by deleting full-stops in a term
- Alternative is to do asymmetric expansion:
  - Enter: ***window*** Search: ***window, windows***
  - Enter: ***windows*** Search: ***Windows, windows, window***
  - Enter: ***Windows*** Search: ***Windows***
- Potentially more powerful, but less efficient

Normalisation: other languages

- Accents: ***résumé*** vs. ***resume***.
- Most important criterion:
  - How are your users likely to write their queries for these words?

- Even in languages that standardly have accents, users often may not type them

- German: ***Tuebingen*** vs. ***Tübingen***
  - Should be equivalent

- <u>**In a practice lab sheet, you will do some preprocessing tasks in python which will help with your coursework.**</u>

OUTLINE

1) What is NLP and where is it used?
2) A typical application: sentiment analysis
3) Properties of natural language and intro to mathematical methods
   3.1) The word level
   3.2) The sentence level
   3.3) Beyond the sentence level for discourse
   3.4) Intro to dialogue and its challenges

Mathematical foundations

- Early work was based on **formal models** of language **often using logical rules, informed by formal linguistics**.
- Now the overwhelmingly most successful methods are **statistical and probabilistic** in nature.
- They may have greater to lesser degrees of ‘linguistic’ information like phrase structure, parts of speech etc.
- *Rules, schmules!* Currently the trend is to have less and less linguists writing rules involved:

> *“Every time I fire a linguist, the performance of the speech recognizer goes up.”*
>
> Fred Jelinek, leading pioneer of modern day automatic speech recognition (ASR)

Mathematical foundations

- However, there's still a use for the classical insights.
- Linguists are still the only ones to point out difficult examples with classical **puzzles of meaning**:

  'Every lecturer gave a student a 1st'

  How many students got 1^{st}'s? One or several?
- And it's still difficult to get an AI system to do proper reasoning without a rule/logic-based **knowledge base**.

  User: 'Book a flight to Denver on Tuesday'

  Sys: 'Okay, where from?'
- But why are the data-driven statistical methods so powerful?

Mathematical foundations

- In a **corpus** of text (or dialogue) you get many regular **patterns**.
- If you understand those patterns systematically, you can figure out what is being talked about, as it's similar to other examples.
- Simple methods can **scale** very fast.
- What are some of these **systematic properties**?

KEY POINT:

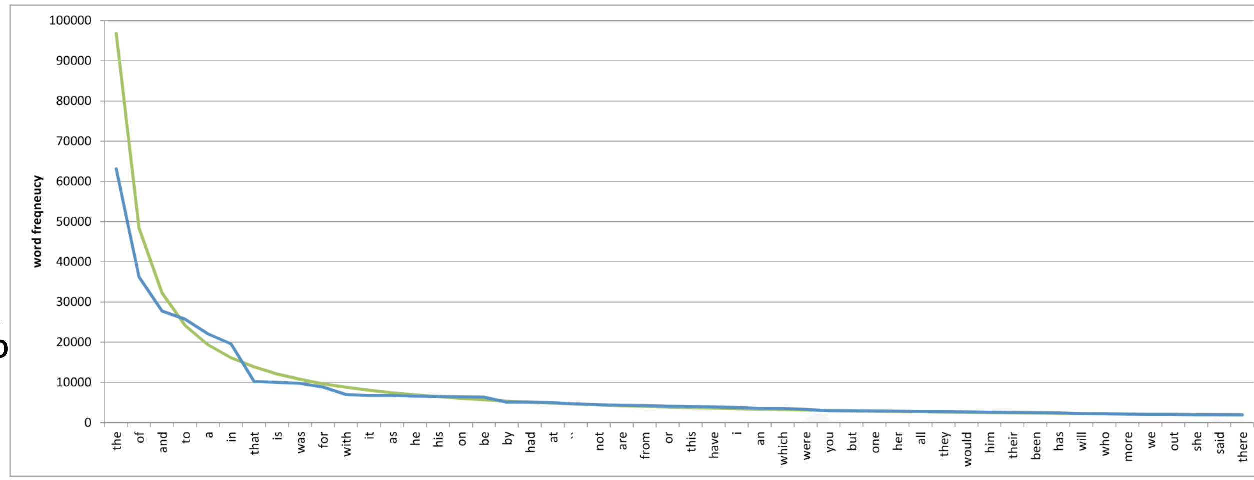
Natural language is Zipfian

Zipf's Law

- The frequency *f* of any word is approximately **inversely proportional** to its rank *r* in the frequency table.

$$f \propto \frac{1}{r}$$

- Brown corpus:
  - rank 1 'the': 7%
  - rank 2 'of': 3.5%
  - rank 3 'and': 2.9%



- This means:
  - We can capture most of the data easily in frequent words.
  - But there is a **very long tail** - almost all words are **rare**.
  - And however big your corpus …
  - ... you will see new words as soon as you look outside it! (***hapax legomenon*** = *word that only occurs once*).

KEY POINT:

Words are not independent

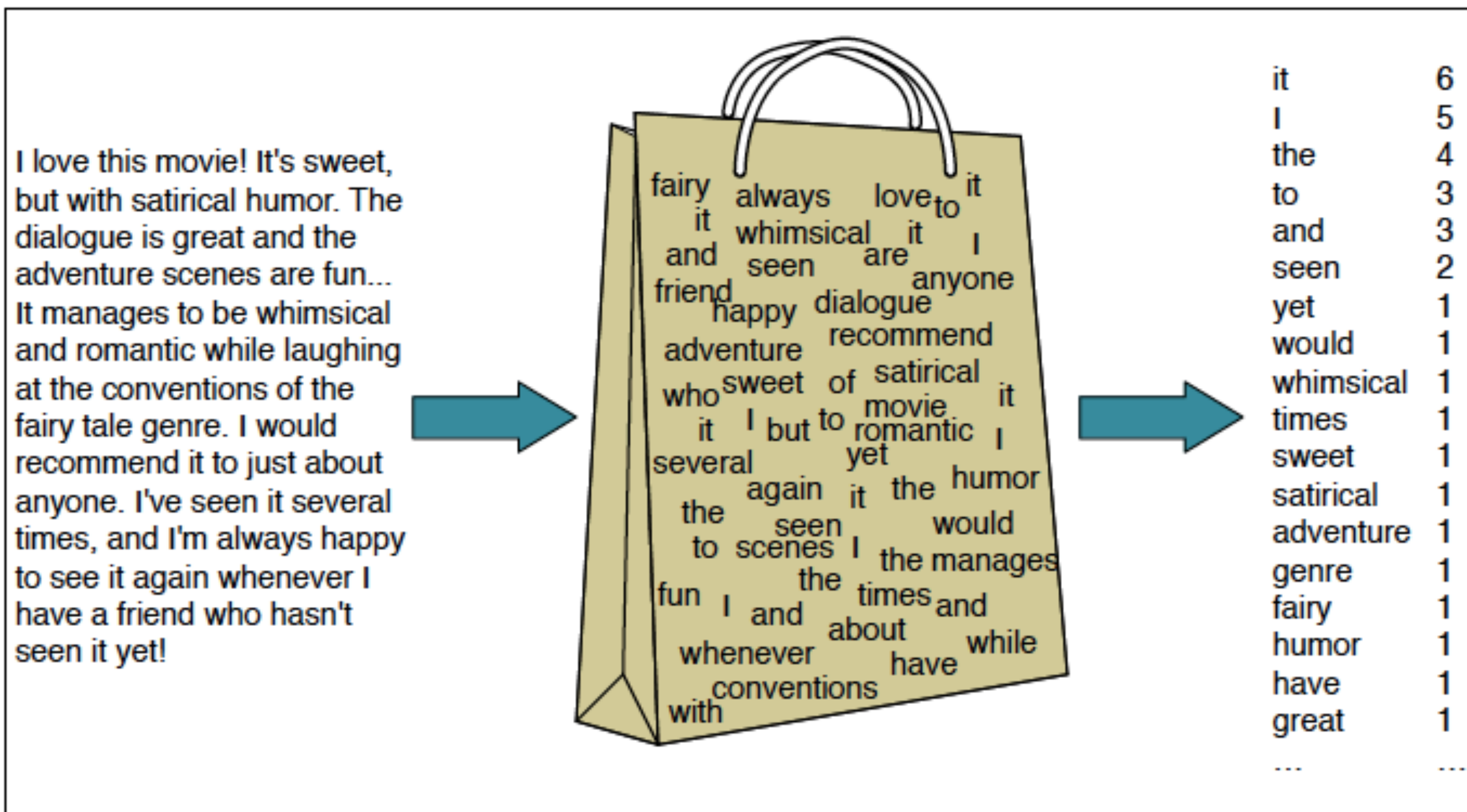
Sentiment Analysis revisited (with Statistical Models)

- How do we model the interaction between words, given single words in a text don’t determine the sentiment on their own?

```
i love @justinbieber #sarcasm
```

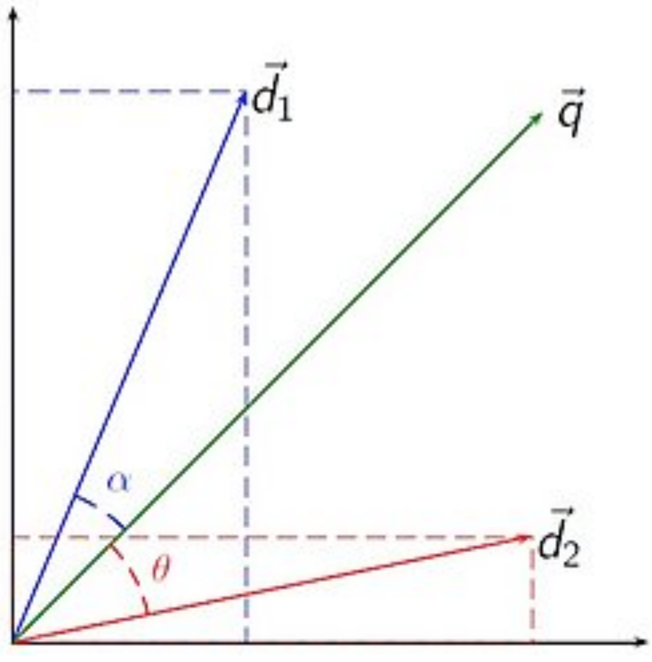
Texts as Feature Spaces

- The simplest way to model interaction between words in texts is characterising a text in terms of the words contained in it.
- **“Bag-of-words” (BoW)** model.



Texts as Feature Spaces

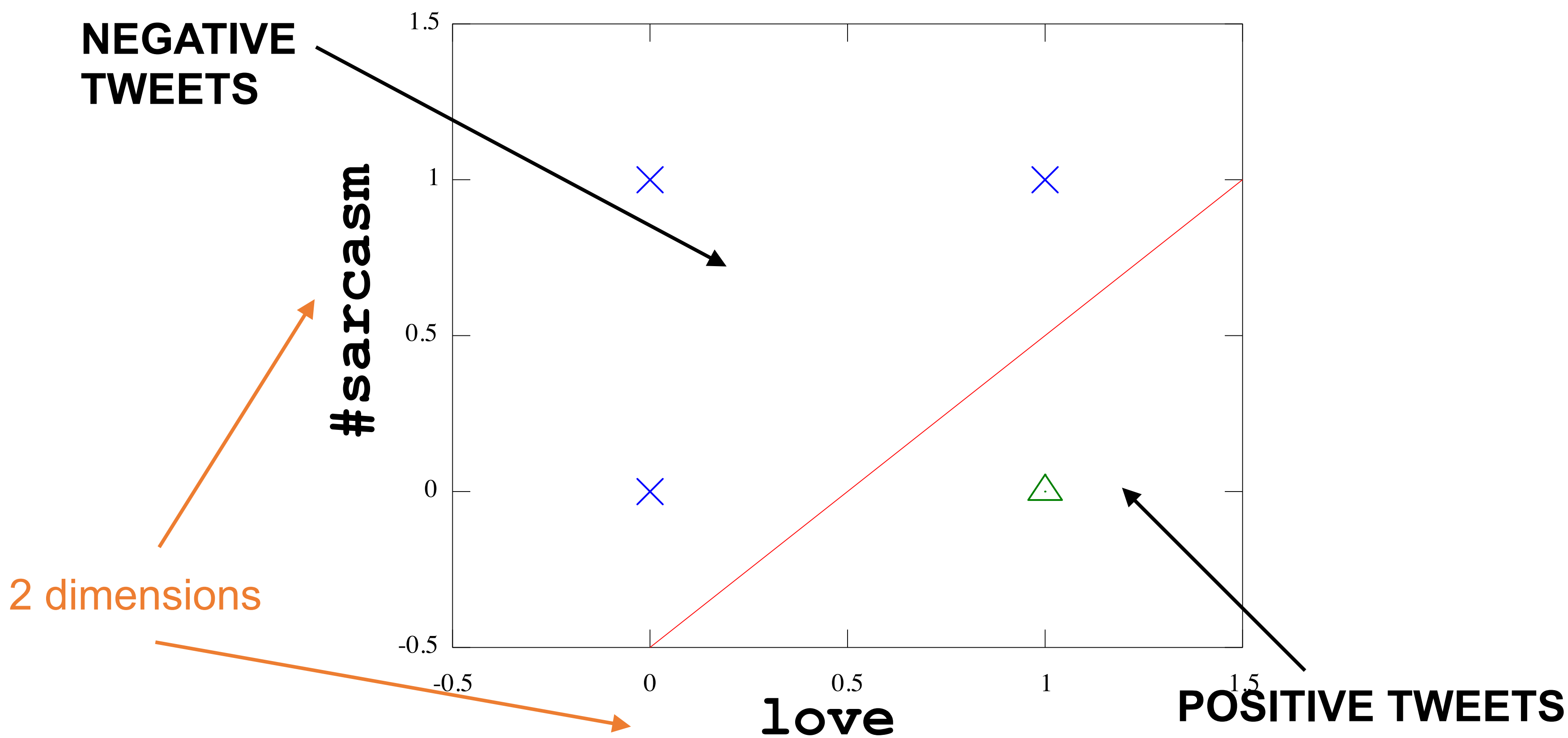
- **Vector space models** (of BoW)
  - words = dimensions
  - documents = vectors



Sentiment Analysis: Data-Driven Classification

- Geometric methods for classification using **Machine Learning**- fit a **class boundary 'line'** in multi-dimensional space using data.

i love @justinbieber #sarcasm



OUTLINE

1) What is NLP and where is it used?
2) A typical application: sentiment analysis
3) Properties of natural language and intro to mathematical methods
   3.1) The word level
   3.2) The sentence level
   3.3) Beyond the sentence level for discourse
   3.4) Intro to dialogue and its challenges

What about …

Milk is good and not expensive

Milk is expensive and not good

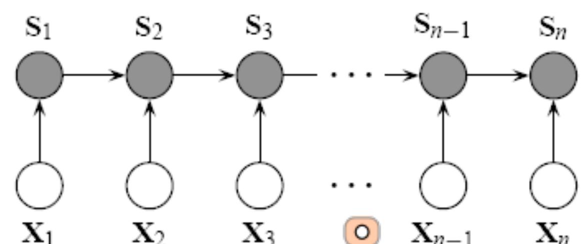
- According to a bag of words model, these two sentences are the same…

KEY POINT:

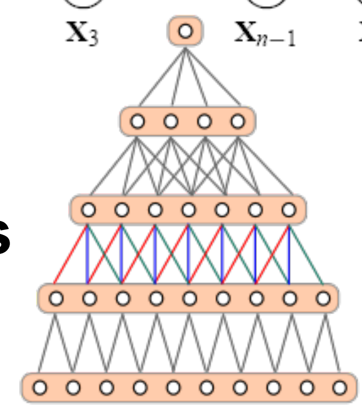
Language is not just a bag of words - order matters

Sequence modelling

- We can get a long way by using **sequences** of words
  - **N-grams (Shannon, 1948)**
    - [milk is], [is good], [good and], [and not], [not expensive]
    - [milk is], [is expensive], [expensive and], [and not], [not good]
  - **Sequence models**
    - **Hidden Markov Models**
    - **Conditional random fields**



  - **Convolutional / recurrent neural nets**



What about …

- Milk is not very good
- Milk is not really very good
- Milk is not bad but good
- As bad as milk is, good things can come from it

- I hate happy birthdays and fluffy clouds
- I love disaster movies

- I like milk
- I like dairy products

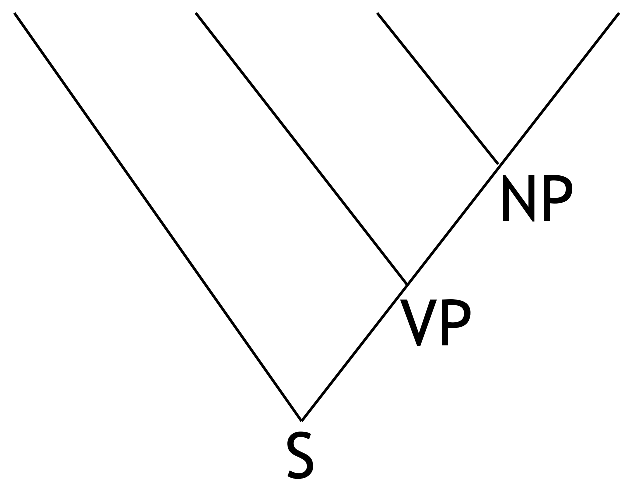
KEY POINT:

Language has hierarchical structure and smaller units compose together into bigger units

Levels of language interpretation

words:	Mary	hires	a	detective	
parts of speech:	PN	VBZ	DET	CN	TAGGING
lemmata:	mary	hire	a	detective	STEMMING

syntax:



PARSING

semantics: ∃x.detective(x) & hire(mary,x)

SEMANTIC PARSING

discourse:

e,x	
hire(e)	detective(x)
subj(e,mary)	obj(e,x)

DISCOURSE PARSING

Syntax: Parsing with Generative Grammars

A Generative System

Each rule has a left-hand symbol
Which ‘generates’ the right-hand side

Sentence	**S** $\longrightarrow$	**NP VP**
Verb Phrase	**VP** $\longrightarrow$	**itV, tV NP**
Transitive Verb	**tV** $\longrightarrow$	**drink, eat**
Intransitive Verb	**itV** $\longrightarrow$	**fly, sleep**
Noun Phrase	**NP** $\longrightarrow$	**vampire, butterfly, blood**

Syntax: Parsing with Generative Grammars

Butterflies sleep

S → NP VP

NP → butterflies

VP → itV

itV → sleep

Syntax: Parsing with Generative Grammars

Vampires drink blood

S $\longrightarrow$ Vampires VP

VP $\longrightarrow$ tV NP

tV $\longrightarrow$ drink

NP $\longrightarrow$ blood

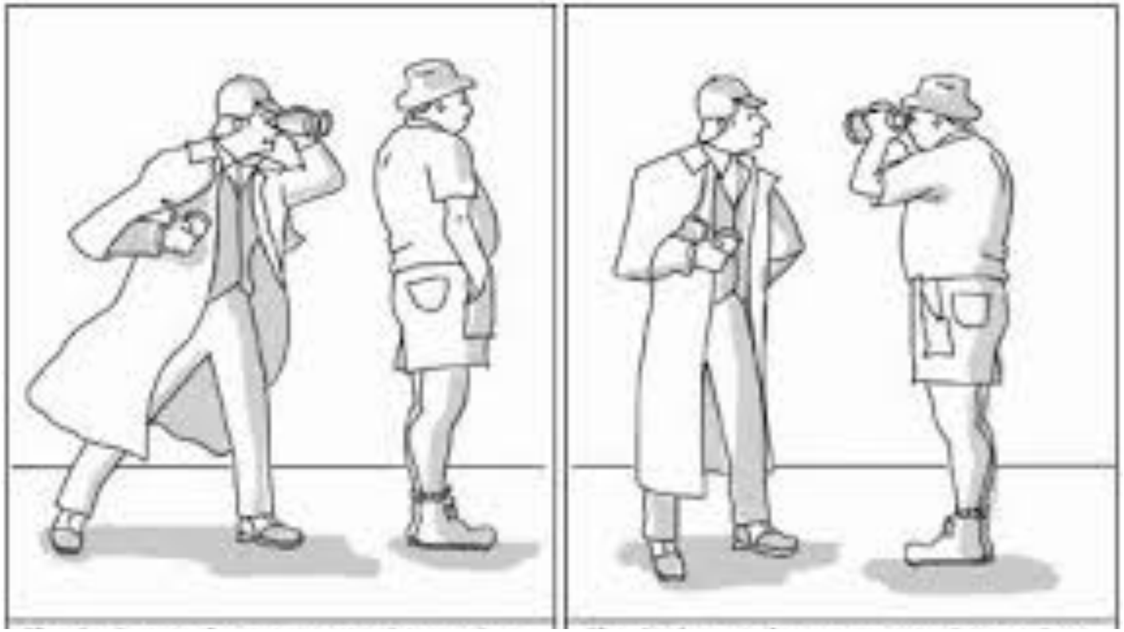
KEY POINT:

Natural language is
ambiguous
and
depends on context

Ambiguity

Syntactic Ambiguity
giving rise to Meaning (Semantic) Ambiguity

Sherlock saw a man with binoculars



Syntactic Ambiguity with Generative Grammars

Sherlock saw a man with binoculars.

S ⟶ Sherlock VP

VP ⟶ saw a man with binoculars

tV ⟶ saw

NP ⟶ a man with binoculars

Meaning 1

Syntactic Ambiguity with Generative Grammars

Sherlock saw a man with binoculars.

S → Sherlock VP PP

VP → saw a man

tV → saw

NP → a man

PP → with binoculars

Meaning 2

Prepositional Phrase

Syntax: Parsing with Logical Grammars

A Logical System

Division and Multiplication

itV: $\frac{S}{NP}$

tV: $\frac{\frac{S}{NP}}{NP}$

itV: fly, sleep

tV: drink, eat

NP: vampire, butterfly, blood

Syntax: Parsing with Logical Grammars

Butterflies sleep.

$$NP \times \frac{S}{NP} = S$$

Vampires drink blood.

$$NP \times \frac{\frac{S}{NP}}{NP} \times NP = NP \times \frac{S}{NP} = S$$

Ambiguity

Semantic/Lexical Ambiguity

Fisher men cast their nets.

The moon cast its light.

Ambiguity

- **<u>How can we deal with the ambiguity of the meaning of a word like ‘cast’ (lexical ambiguity)?</u>**

- **<u>How do we deal with word meaning in general?</u>**
  - **Semantics**
    - **Formal logical methods**- each word maps to a formula
      - ‘cast’ -> *cast_net*
      - ‘cast’ -> *cast_shine*
    - **Distributional methods**- a word’s meaning is defined by its use (where it occurs in a text relative to others)

Guess the missing word

It is difficult to make a single, definitive description of the **folkloric** [redacted] though there are several elements common to many European **legends**. [redacted] were usually reported as bloated in appearance, and **ruddy**, **purplish**, or dark in colour; these characteristics were often attributed to the drinking of **blood**. [· · ·] Indeed, **blood** was often seen seeping from the mouth and nose of the [redacted] when it was seen in its **shroud** or **coffin** and its left eye was often open. [· · ·] In Christianity, the [redacted] was viewed as ”a **dead** person who retained a semblance of life and could leave its **grave**-much in the same way that Jesus had risen after his **death** and **burial** and appeared before his followers. In Asia, [· · ·] a [redacted] wanders around animating **dead bodies** at night, attacking the living much like a **ghoul**.

It is difficult to make a single, definitive description of the **folkloric** **vampire**, though there are several elements common to many European **legends**. **Vampire**s were usually reported as bloated in appearance, and **ruddy**, **purplish**, or dark in colour; these characteristics were often attributed to the drinking of **blood**. $[\cdots]$ Indeed, **blood** was often seen seeping from the mouth and nose of the **vampire** when it was seen in its **shroud** or **coffin** and its left eye was often open. $[\cdots]$ In Christianity, the **vampire** was viewed as ”a **dead** person who retained a semblance of life and could leave its **grave**-much in the same way that Jesus had risen after his **death** and **burial** and appeared before his followers. In Asia, $[\cdots]$ a **vampire** wanders around animating **dead bodies** at night, attacking the living much like a **ghoul**.

Guess the missing word

[redacted] are beautiful, flying insects with large scaly wings. Like all insects, they have six jointed legs, 3 body parts, a pair of antennae, compound eyes, and an exoskeleton. The three body parts are the head, thorax (the chest), and abdomen (the tail end). The [redacted]'s body is covered by tiny sensory hairs. The four wings and the six legs of the [redacted] are attached to the thorax. The thorax contains the muscles that make the legs and wings move. [redacted] are very good fliers. They have two pairs of large wings covered with colorful, iridescent scales in overlapping rows. Lepidoptera ([redacted] and moths) are the only insects that have scaly wings. The wings are attached to the [redacted]'s thorax (mid-section). Veins support the delicate wings and nourish them with blood.

Butterflies are beautiful, flying insects with large scaly wings. Like all insects, they have six jointed legs, 3 body parts, a pair of antennae, compound eyes, and an exoskeleton. The three body parts are the head, thorax (the chest), and abdomen (the tail end). The butterfly's body is covered by tiny sensory hairs. The four wings and the six legs of the butterfly are attached to the thorax. The thorax contains the muscles that make the legs and wings move. Butterflies are very good fliers. They have two pairs of large wings covered with colorful, iridescent scales in overlapping rows. Lepidoptera (butterflies and moths) are the only insects that have scaly wings. The wings are attached to the butterfly's thorax (mid-section). Veins support the delicate wings and nourish them with blood.

Distributional Hypothesis

- Words that occur in similar contexts tend to have similar meanings. This insight was first formulated by Harris (1954) who said:
  - "oculist and eye-doctor . . . occur in almost the same environments"
- and more generally that
  - "If A and B have almost identical environments. . . we say that they are synonyms."

- The most famous statement of the principle comes a few years later from the linguist Firth (1957), who phrased it as

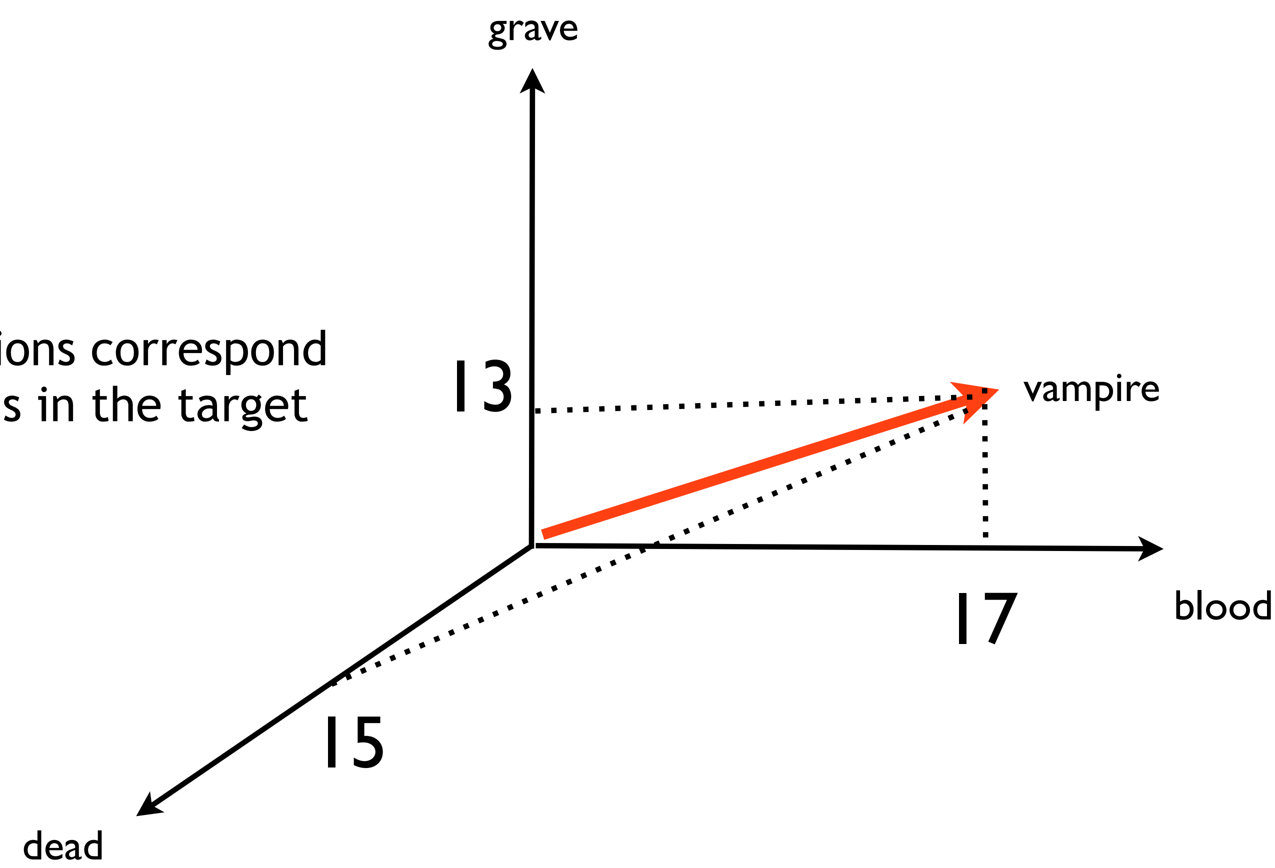
  - "You shall know a word by the company it keeps!"

- The meaning of a word is thus related to the **distribution** of the words around it.

Distributional Meaning: Words as vectors

$$\overrightarrow{\textit{vampire}} = (17, 13, 15)$$

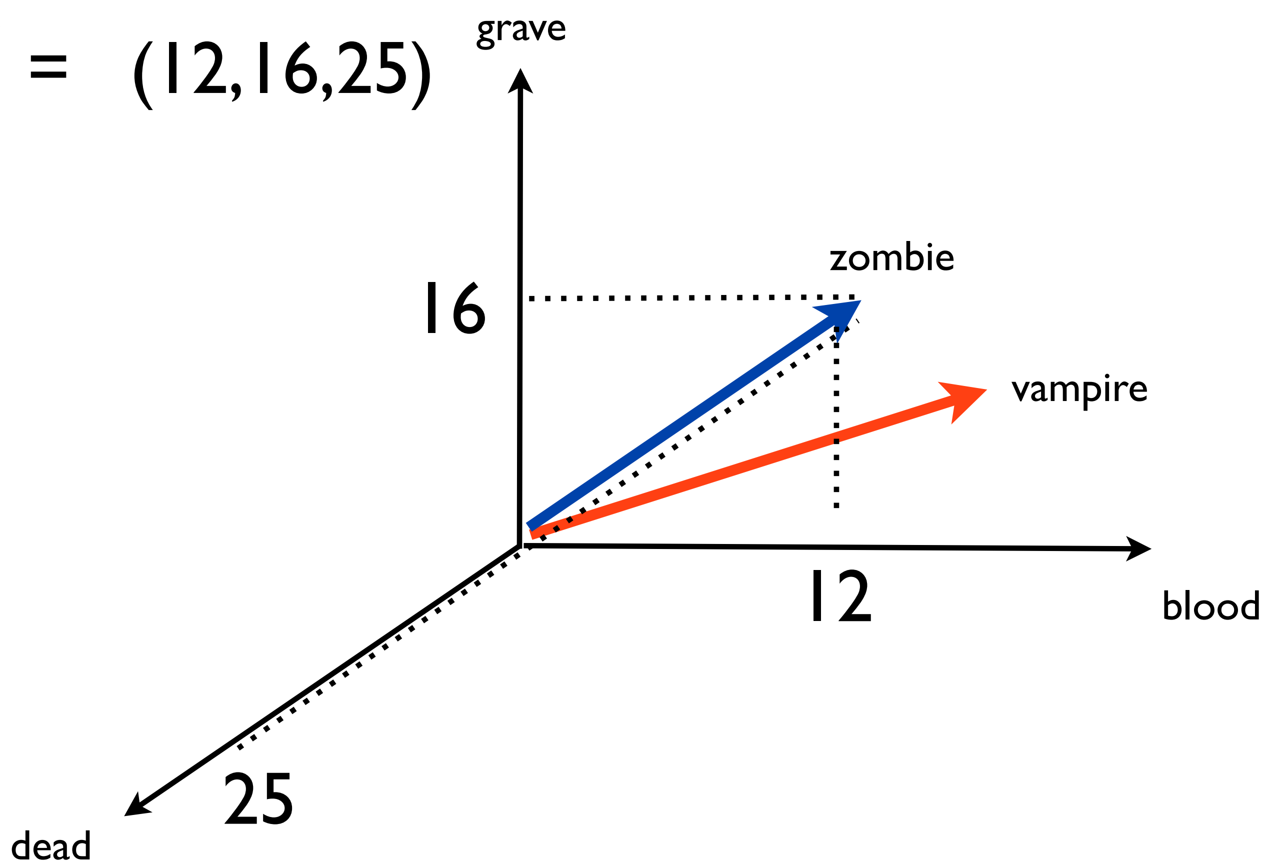
Different dimensions correspond to different words in the target word's context



Words as vectors

$$\overrightarrow{vampire} = (17,13,15)$$

$$\overrightarrow{zombie} = (12,16,25)$$

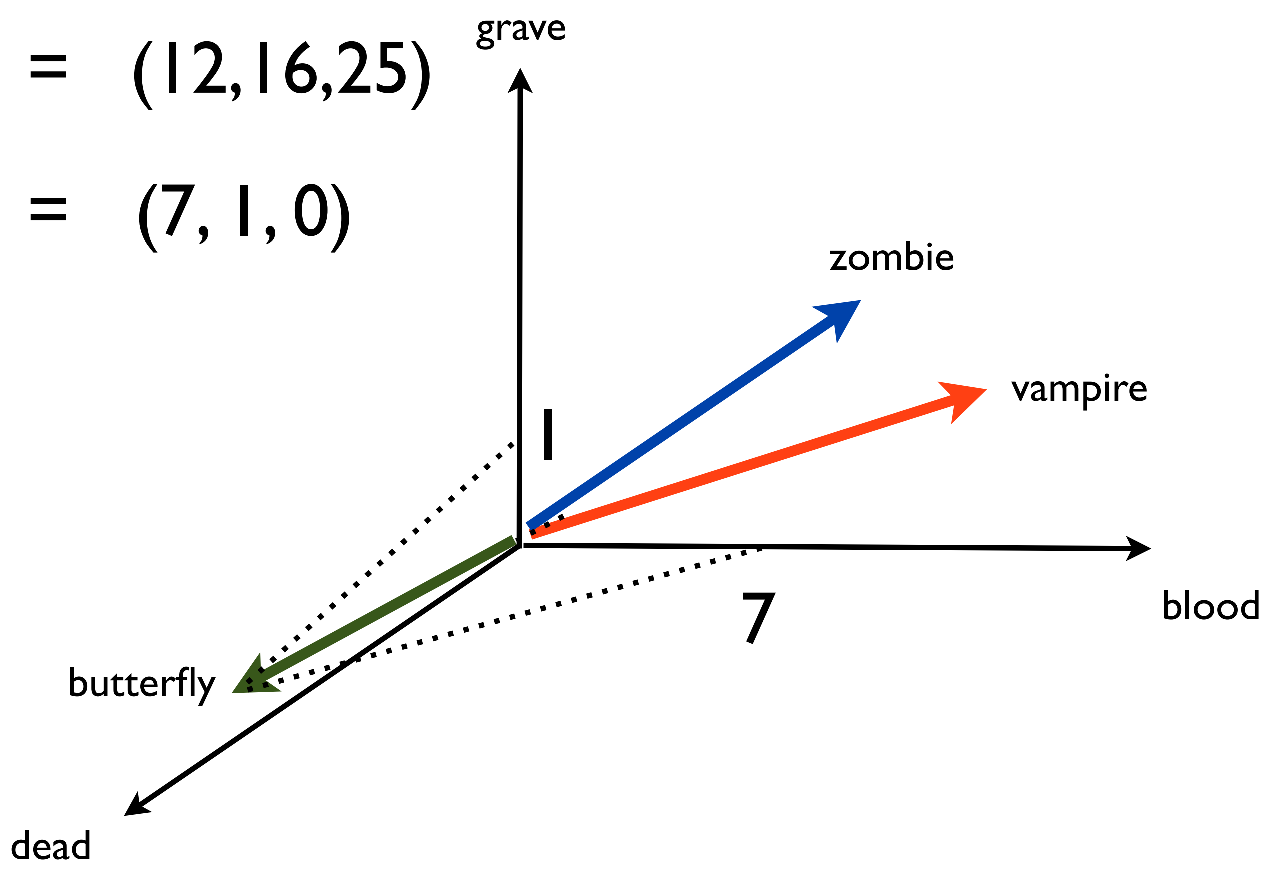


Words as vectors

$\overrightarrow{vampire} = (17,13,15)$

$\overrightarrow{zombie} = (12,16,25)$

$\overrightarrow{butterfly} = (7, 1, 0)$



OUTLINE

1) What is NLP and where is it used?
2) A typical application: sentiment analysis
3) Properties of natural language and intro to mathematical methods
   3.1) The word level
   3.2) The sentence level
   3.3) Beyond the sentence level for discourse
   3.4) Intro to dialogue and its challenges

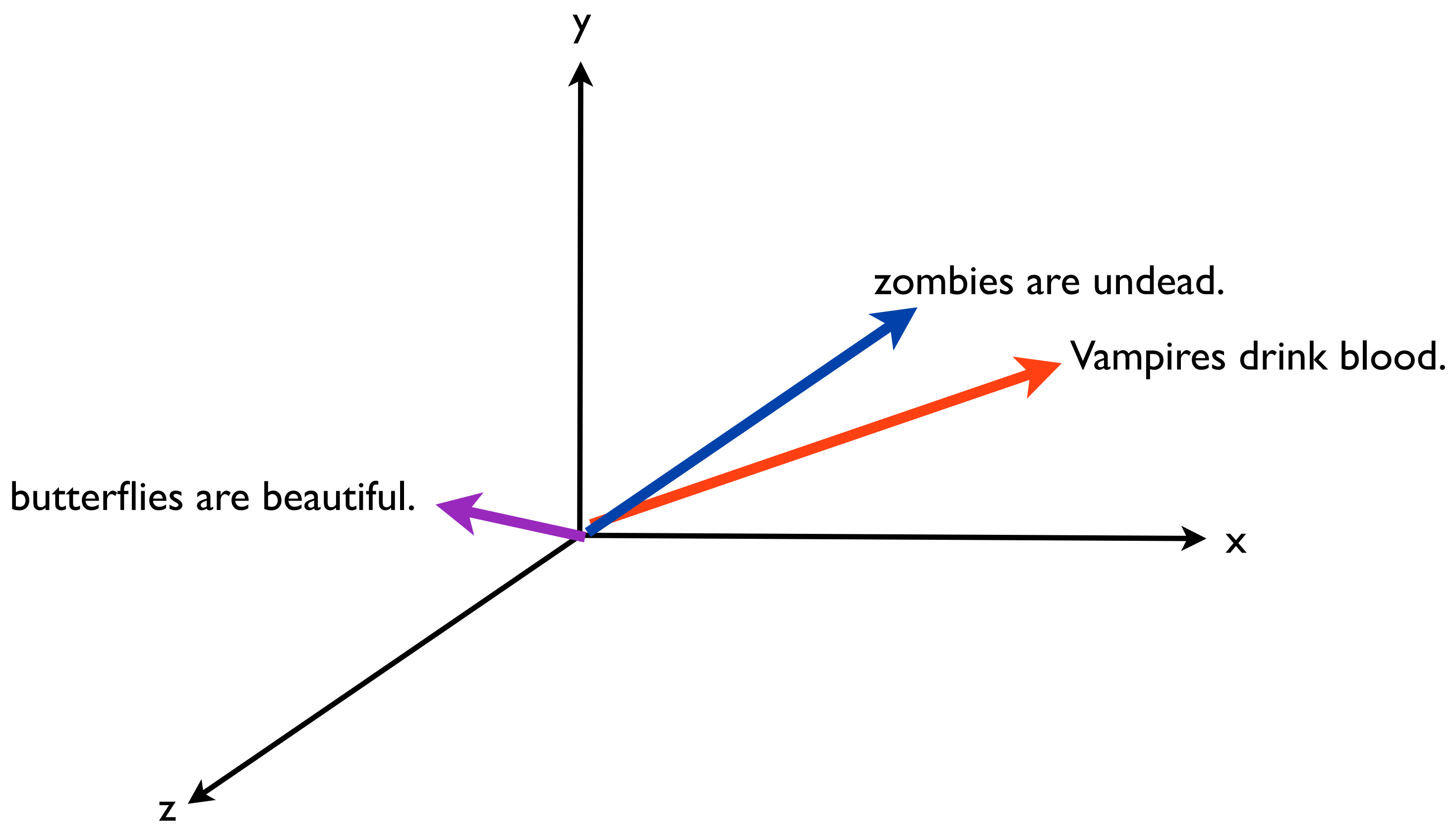
Guess the missing sentence

[redacted] **Vampire** were usually reported as bloated in appearance, and **ruddy**, **purplish**, or dark in colour; these characteristics were often attributed to the drinking of **blood**. [· · ·] Indeed, **blood** was often seen seeping from the mouth and nose of the **vampire** when it was seen in its **shroud** or **coffin** and its left eye was often open. [· · ·] In Christianity, the **vampire** was viewed as ”a **dead** person who retained a semblance of life and could leave its **grave**-much in the same way that Jesus had risen after his **death** and **burial** and appeared before his followers. In Asia, [· · ·] a **vampire** wanders around animating **dead bodies** at night, attacking the living much like a **ghoul**.

Distributional Hypothesis

- Can we update Firth (1957) to the following?
  - "You shall know a *sentence* by the company it keeps!"
- The meaning of a sentence is thus related to the **distribution of the sentences** around it.
- However, the problem is in the **sparsity** of sentences- the steepness of the curve of the Zipfian distribution of sentences is **far steeper** than that for words.

Sentences as vectors?



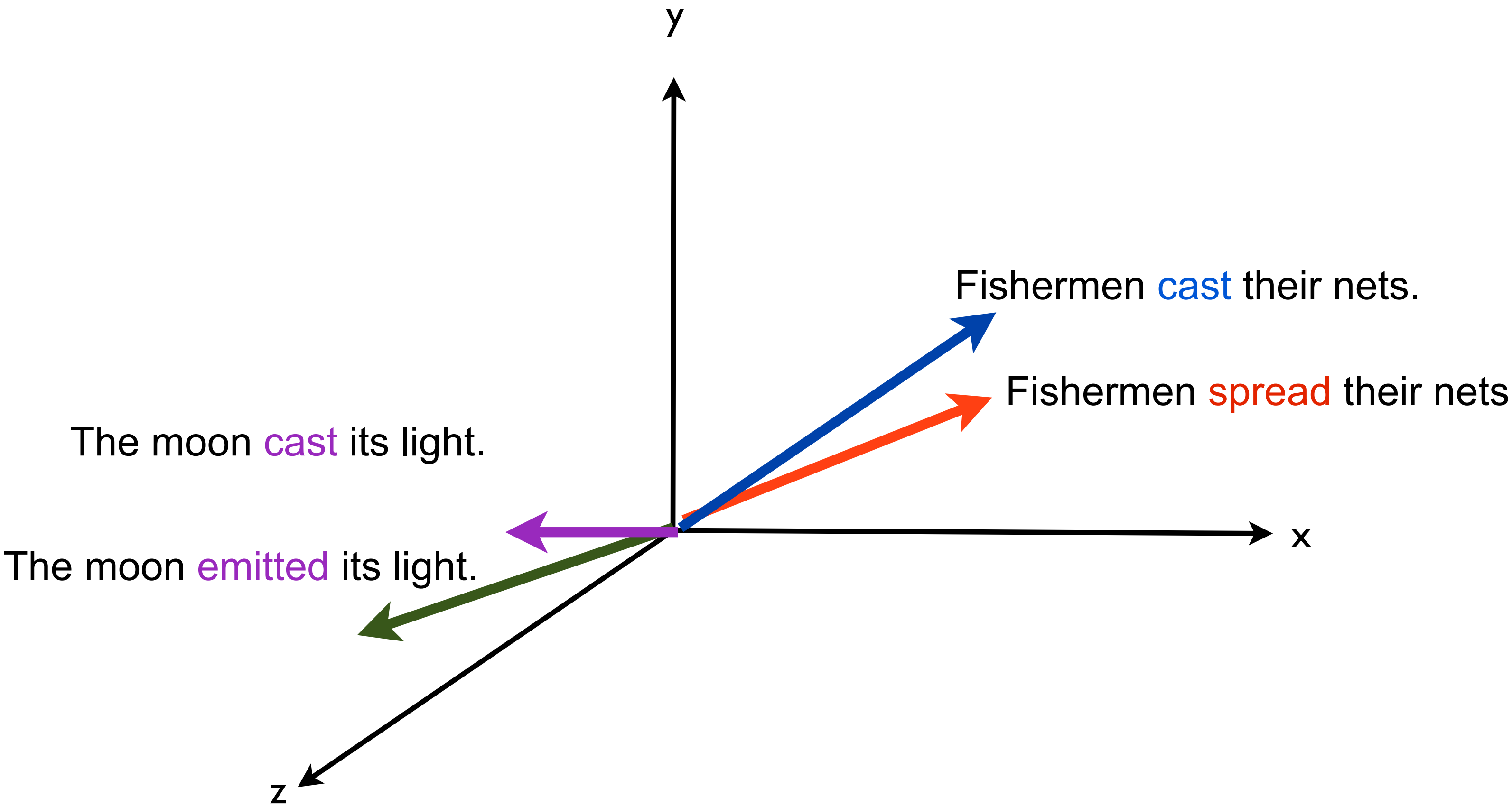
one way to avoid sparsity: simple vector operations

$$\overrightarrow{\text{vampires kill men}} = \overrightarrow{\text{vampires}} + \overrightarrow{\text{kill}} + \overrightarrow{\text{men}}$$

$$= \overrightarrow{\text{vampires}} \odot \overrightarrow{\text{kill}} \odot \overrightarrow{\text{men}}$$

$$\overrightarrow{\text{vampires kill men}} = \overrightarrow{\text{men kill vampires}}$$

Word Sense Disambiguation



What about words without obvious lexical meaning without context?

- How about words like “he”, and “she” and “it’?
- On their own they don’t mean much, so we have to use context in a way **beyond the sentence** they are in to get their meaning.
- They get their meaning from the **discourse,** i.e. the context where they are used.

OUTLINE

1) What is NLP and where is it used?
2) A typical application: sentiment analysis
3) Properties of natural language and intro to mathematical methods
   3.1) The word level
   3.2) The sentence level
   3.3) Beyond the sentence level for discourse
   3.4) Intro to dialogue and its challenges

What about when language is used by multiple people in a dialogue…

A: I like all milk, which is white and tasty

B: I agree!

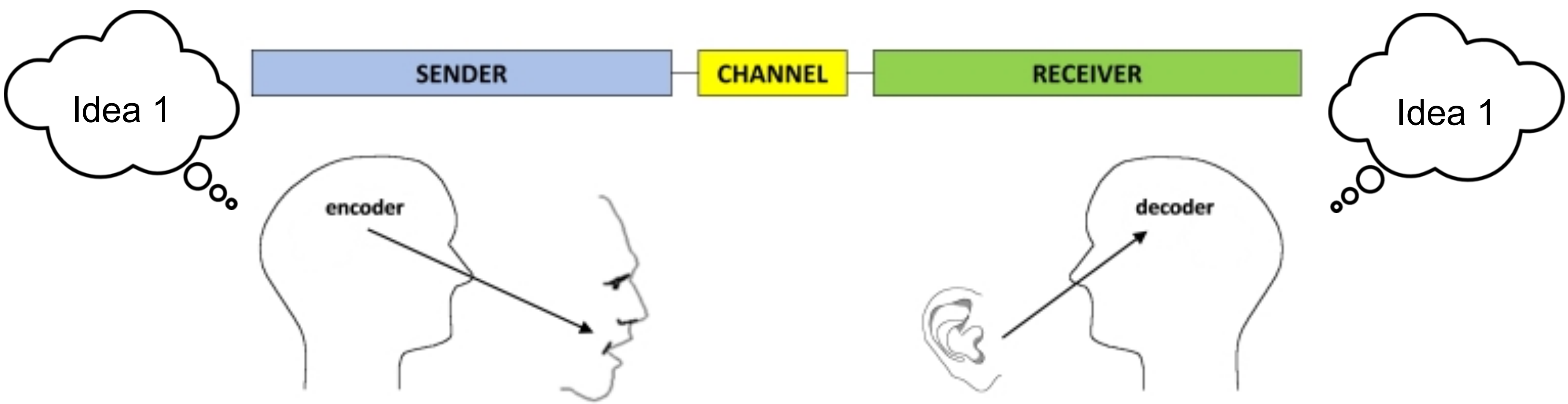
C: No way.

- How can we tell what B and C **mean**?
- How would we design a program which could understand B and C's contributions?

How do people communicate with natural language?

- First models similar to encoder/decoder model (Shannon, 1948).

- Communication based on a common code.



How do people communicate with natural language?

- What about the **missing** words/parts of the linguistic ‘code’- how could a machine compute meaning like a human does?

A: I like all milk, which is white and tasty

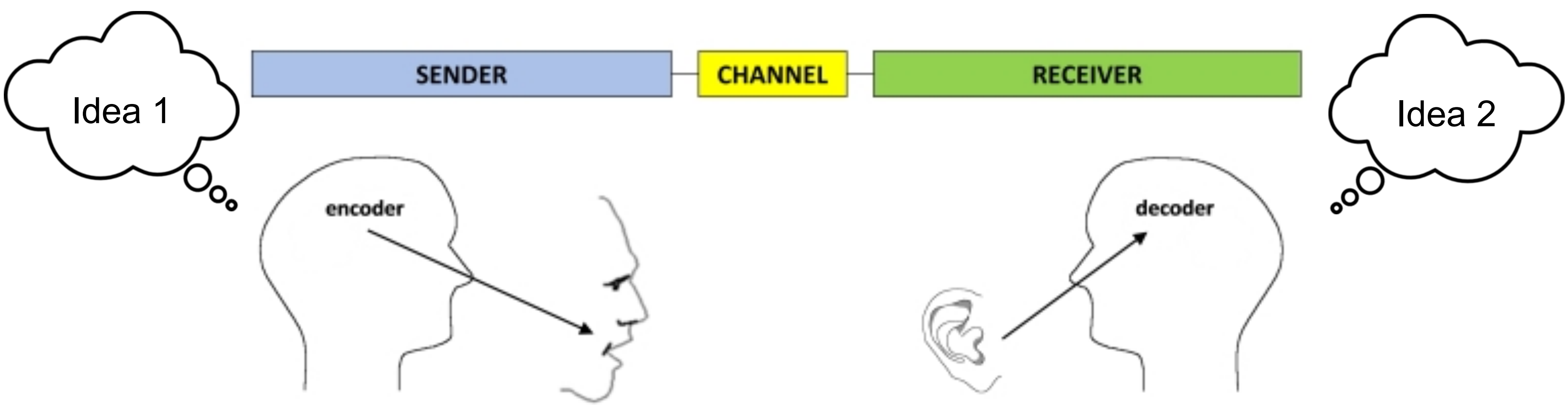
B: I agree!

C: No way.

- B’s turn is missing something and not the complete version: e.g. “I agree that milk is white and tasty”
- Understanding C’s turn relies on understanding B’s turn, e.g. “I disagree that milk is white and tasty”

How can people *mis*communicate?

- Just noise in signal? More recent theories about aligning internal representations via ***communicative grounding*** (Clark 1996) mechanisms.
- A. ‘Put the apple over there’
  - B. ‘Where did you mean?’ (clarification)
  - A. ‘No, in the corner’ (repair)



How can people *mis*communicate?

- Self-repair/disfluency (every 25 words of natural dialogue), but not taken seriously by engineers:

  *“But one of **the, the** two things that I’m really. . .”*

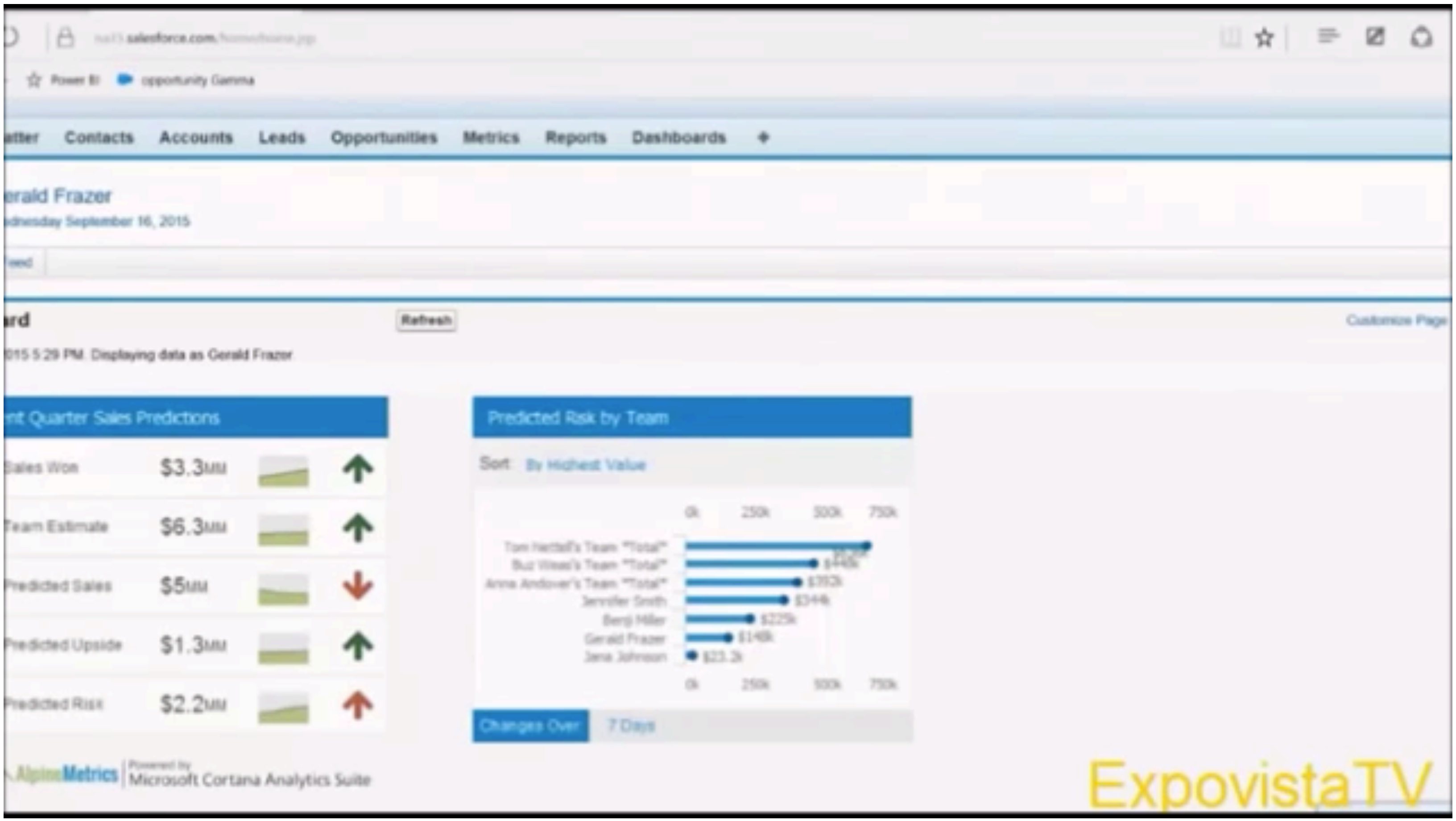
  *“Our situation is **just a little bit, kind of** the opposite of that”*

  *“and you know it’s like **you’re, I mean,** employments are contractual by nature anyway”*

KEY POINT:

Dialogue is challenging and messy and requires context!

And hard for systems…



How do we build systems to speak with humans?

- Dialogue system designers struggle to deal with the rich range of human dialogue behaviour and what people **mean** in their utterances/texts.
- However, many useful systems use simple assumptions to get things working.

How do we build systems to speak with humans?

- Google Dialogflow uses breaks things down to **intents** and **entities** and context variables.
- An intent is the recognized meaning of the user’s intention e.g. I want a pizza -> *#orderfood*
- An entity is an individuated thing e.g. I want a pizza -> *entity:food=pizza*
- **<u>As a practice exercise you will build a simple Google Dialogflow chatbot.</u>**

Reading

- Christopher D. Manning and Hinrich Schuetze (2003/1999). **Foundations of Statistical Natural Language Processing.** Chapter 1

- (optional) If you aren't familiar with Python / you're getting started with natural language data and corpora:
  - **Python tutorial** (online) https://docs.python.org/3/tutorial/

  - **NLTK book** (online), Chapters 1 and 2 https://www.nltk.org/book/