Natural Language Processing 1

Live Q & A: Semantics

Katia Shutova

ILLC
University of Amsterdam

25 November 2020

Characteristic model

- Weights given to the vector components express how *characteristic* a given context is for word w.
- Pointwise Mutual Information (PMI)

$$PMI(w,c) = \log \frac{P(w,c)}{P(w)P(c)} = \log \frac{P(w)P(c|w)}{P(w)P(c)} = \log \frac{P(c|w)}{P(c)}$$

$$P(c) = \frac{f(c)}{\sum_k f(c_k)}, \qquad P(c|w) = \frac{f(w,c)}{f(w)},$$

$$PMI(w,c) = \log \frac{f(w,c)\sum_k f(c_k)}{f(w)f(c)}$$

$f(w,c)$: frequency of word w in context c
$f(w)$: frequency of word w in all contexts
$f(c)$: frequency of context c

Our reference text

Douglas Adams, *Mostly harmless*

The major difference between a thing that might go wrong and a thing that cannot possibly go wrong is that when a thing that cannot possibly go wrong goes wrong it usually turns out to be impossible to get at or repair.

- **Example:** Produce distributions using a word window, PMI-based model

The semantic space

Douglas Adams, *Mostly harmless*

The major difference between a thing that might go wrong and a thing that cannot possibly go wrong is that when a thing that cannot possibly go wrong goes wrong it usually turns out to be impossible to get at or repair.

- Assume only keep open-class words.
- **Dimensions:**

difference	impossible	thing
get	major	turns
go	possibly	usually
goes	repair	wrong

Frequency counts...

Douglas Adams, *Mostly harmless*

The major difference between a thing that might go wrong and a thing that cannot possibly go wrong is that when a thing that cannot possibly go wrong goes wrong it usually turns out to be impossible to get at or repair.

- **Counts:**

difference 1	impossible 1	thing 3
get 1	major 1	turns 1
go 3	possibly 2	usually 1
goes 1	repair 1	wrong 4

Conversion into 5-word windows...

Douglas Adams, *Mostly harmless*

The major difference between a thing that might go wrong and a thing that cannot possibly go wrong is that when a thing that cannot possibly go wrong goes wrong it usually turns out to be impossible to get at or repair.

- $\emptyset$ $\emptyset$ **the** major difference
- $\emptyset$ the **major** difference between
- the major **difference** between a
- major difference **between** a thing
- ...

Distribution for *wrong*

Douglas Adams, *Mostly harmless*

The major difference between a thing that [might go wrong and a] thing that cannot [possibly go wrong is that] when a thing that cannot [possibly go [wrong goes wrong] it usually] turns out to be impossible to get at or repair.

- **Distribution (frequencies):**

difference 0	impossible 0	thing 0
get 0	major 0	turns 0
go 3	possibly 2	usually 1
goes 2	repair 0	wrong 2

Distribution for *wrong*

- **Overall word frequency counts in the corpus:**

difference 1	impossible 1	thing 3
get 1	major 1	turns 1
go 3	possibly 2	usually 1
goes 1	repair 1	wrong 4

- **Contexts for *wrong* (frequencies):**

difference 0	impossible 0	thing 0
get 0	major 0	turns 0
go 3	possibly 2	usually 1
goes 2	repair 0	wrong 2

- **PMI of *wrong* (w) and *usually* (c)**

$$PMI(w,c) = \log \frac{f(w,c)\sum_k f(c_k)}{f(w)f(c)} = log\frac{1*20}{4*1} = log5 \approx 0.70$$

Distribution for *wrong*

Douglas Adams, *Mostly harmless*

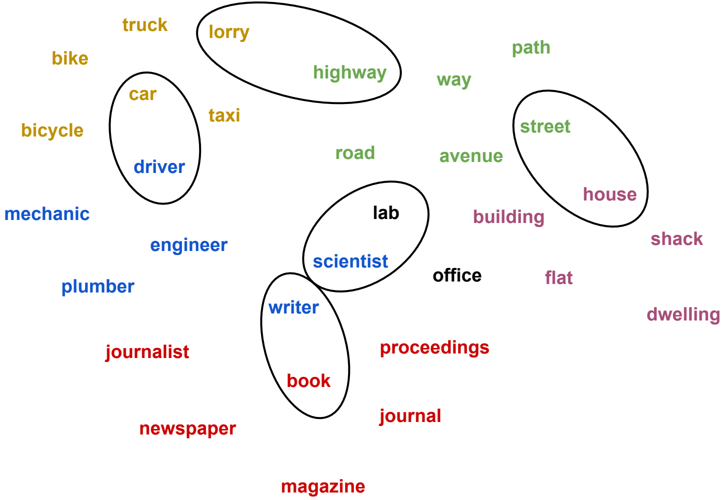
The major difference between a thing that [might go wrong and a] thing that cannot [possibly go wrong is that] when a thing that cannot [possibly go [wrong goes wrong] it usually] turns out to be impossible to get at or repair.

- **Distribution (PPMIs):**

difference 0	impossible 0	thing 0
get 0	major 0	turns 0
go 0.70	possibly 0.70	usually 0.70
goes 1	repair 0	wrong 0.40

Question 1: Clustering nouns

Question 1: Clustering nouns



Question 2: Modelling meaning change

Imagine that you are given a dataset of historical texts (1860-2020). It contains 5 parts each covering a particular time period. This type of data captures how language evolves.

How could you use **distributional semantics** or **word embeddings** to model change in word meaning over time?

Question(s) 3: Discussion

1. How are the following tasks / methods affected by the differences in genre? Which of the tasks is more sensitive to this difference and why?
   - n-gram language modelling
   - PoS tagging
   - probabilistic syntactic parsing
   - distributional semantics and word embeddings
2. For the challenges and applications we discussed in Lecture 1, which problems can the above methods solve?

Question 4: Lexicon vs BOW

Discuss the similarities and differences between the lexicon-based approach to sentiment classification you saw in Practical 1 and the BOW approach of Practical 2.

Acknowledgement

Some slides were adapted from Aurelie Herbelot