Advanced Natural Language Processing

Intro

David Schlangen
University of Potsdam, MSc Cognitive Systems
Winter 2019 / 2020

Welcome

to U of Potsdam, MSc CogSys

Departmental Welcome Reception: Next week Monday (Oct 21), 4pm, 2.14.0.35

You

- Backgrounds: Computational Linguistics, Linguistics, Computer Science, Maths, other?

- Backgrounds: Country of School Degree?

Your Instructor

- David Schlangen
- since 2019 Professor @UP (“Foundations of Computational Linguistics”); 2010-2019 Professor @ Bielefeld U; before: PostDoc in Potsdam & Edinburgh
- PhD from University of Edinburgh; MA in CL & CS from U Bonn

how to address me:

- David
- also works: Prof. Schlangen; Professor; …
- in general good strategy: wait for how you are addressed, and then follow that

today

- what’s NLP? examples, first attempt at formalisation
- NLP and neighbouring fields
- central themes in NLP
- this class, administrativia

“advanced NLP” — what’s that?

- Natural Language Processing is the set of methods for making human language accessible to computers. [Eisenstein 2019, p. 13]

what kinds of NLP applications do you know / (would) find useful?

- ?

some NLP applications
(no particular order)

- Machine Translation
- Grammar Checking
- Spell Checking / “Did you mean?”
- Predictive Text
- Dialogue Systems / “intelligent assistants” / Robots?
- Summarisation
- Information Extraction
- Text Classification: Spam; Sorting by Genre; etc. etc.

formalising a bit…

- NLP is the creation of mappings from NL to R:
  NL → R
  where R is a *representation* that “makes NL accessible to computer”

- and / or back, from R → NL

- more precisely, mappings from / to (or relations between):

  - V, vocabulary of symbols; typically large (|V|)

  - V*, sequences of symbols from vocabulary

  - Y, vocabulary of labels; typically small

  - $\mathscr{Y}$, set of structured objects (e.g. trees)

  - $\mathbb{R}$

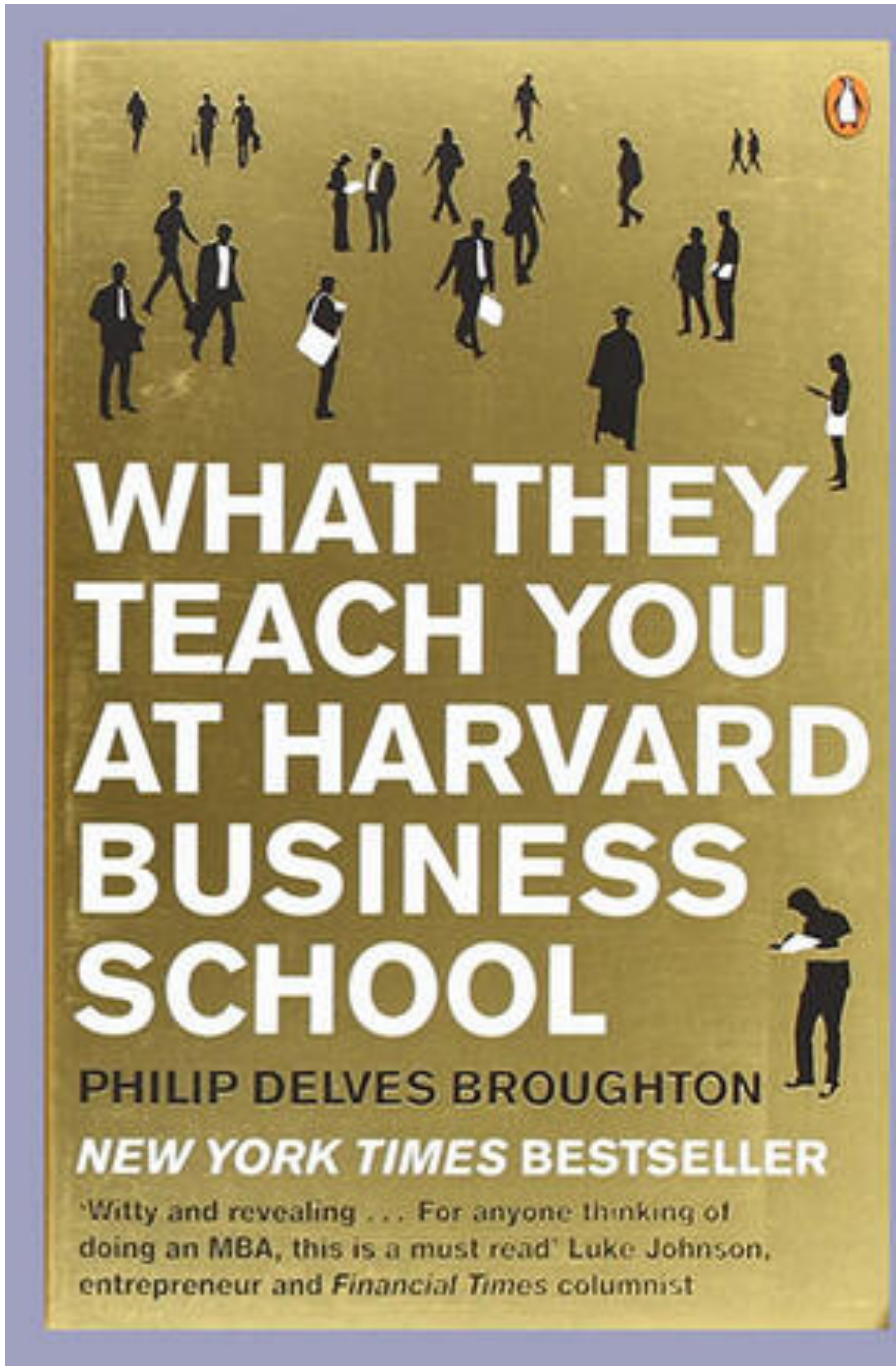
why is this hard?

why not just write a simple program that realises NL → R?

- natural language is massively ambiguous. (You just don’t normally notice, which is interesting.)

DOCTOR WHO TESTED POSITIVE FOR EBOLA
TOOK SUBWAY, WENT BOWLING IN BROOKLYN

If your DOG does a POO Please put it in a litter bin.
Please help keep our open spaces clean.

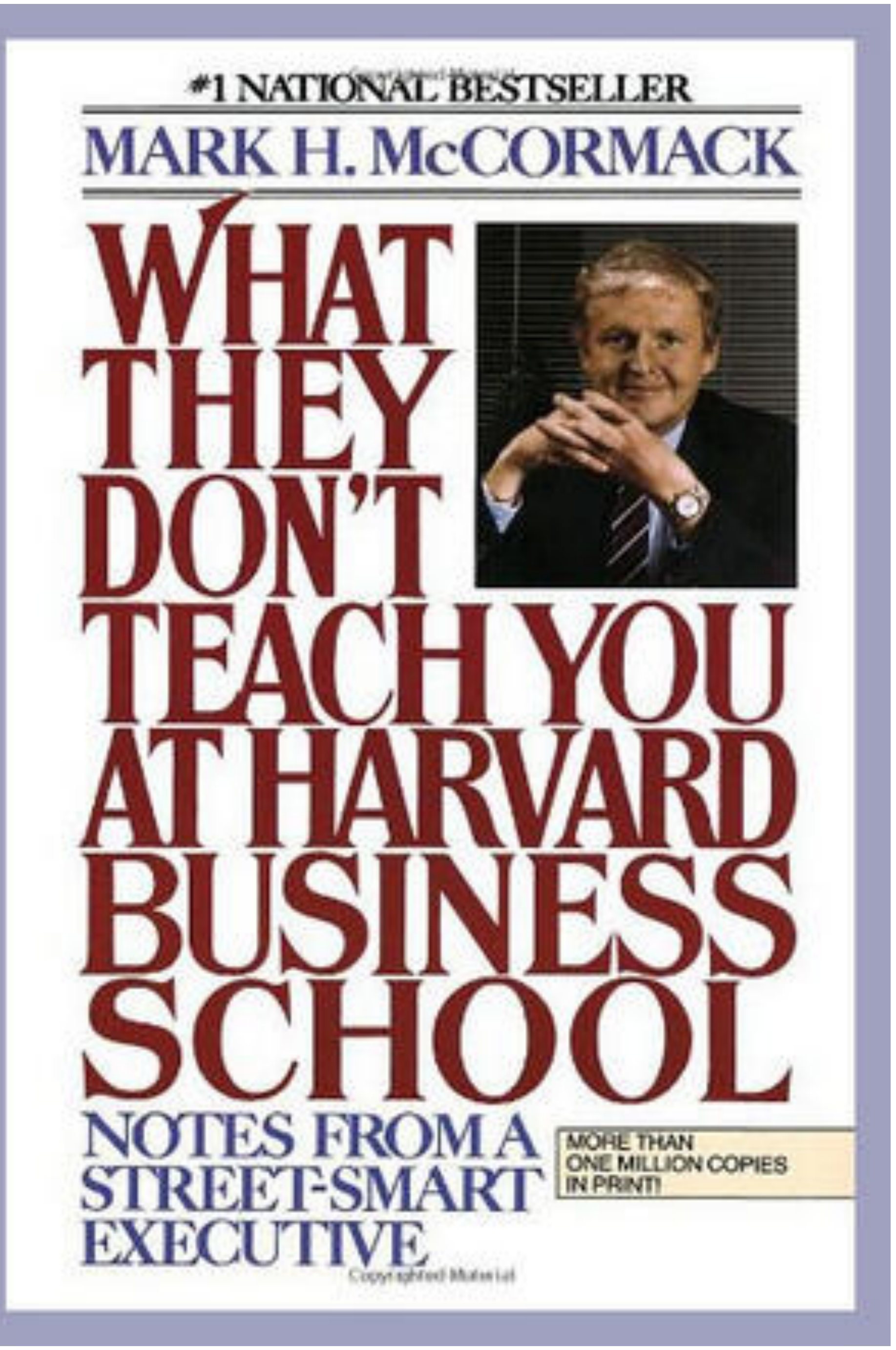


WHAT THEY TEACH YOU AT HARVARD BUSINESS SCHOOL

PHILIP DELVES BROUGHTON

NEW YORK TIMES BESTSELLER

'Witty and revealing . . . For anyone thinking of doing an MBA, this is a must read' Luke Johnson, entrepreneur and *Financial Times* columnist



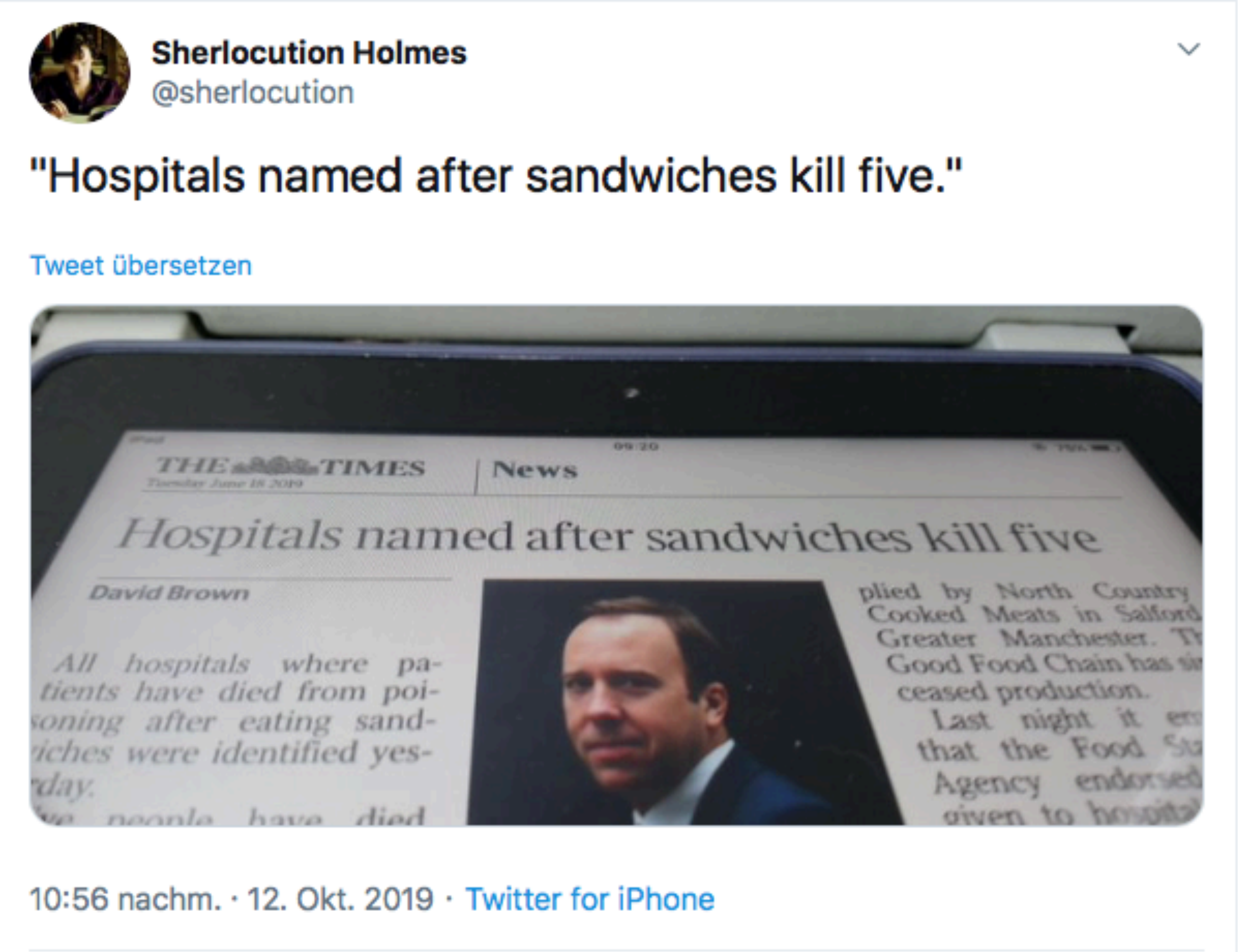
#1 NATIONAL BESTSELLER

MARK H. McCORMACK

WHAT THEY DON'T TEACH YOU AT HARVARD BUSINESS SCHOOL

NOTES FROM A STREET-SMART EXECUTIVE

MORE THAN ONE MILLION COPIES IN PRINT!



Sherlocution Holmes
@sherlocution

"Hospitals named after sandwiches kill five."

Tweet übersetzen

THE TIMES | News

Tuesday June 18 2019

Hospitals named after sandwiches kill five

David Brown

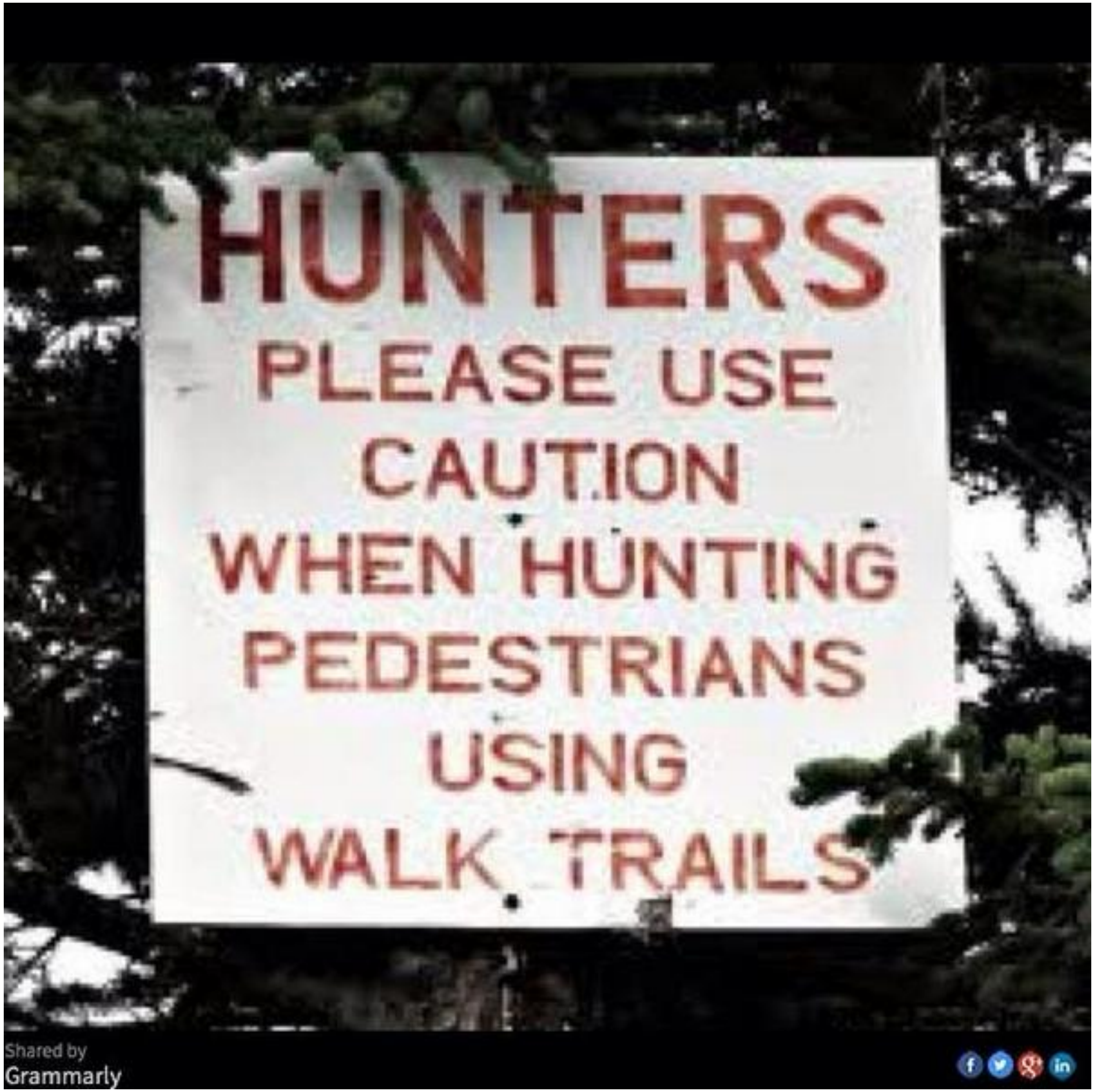
All hospitals where patients have died from poisoning after eating sandwiches were identified yesterday.

ive people have died

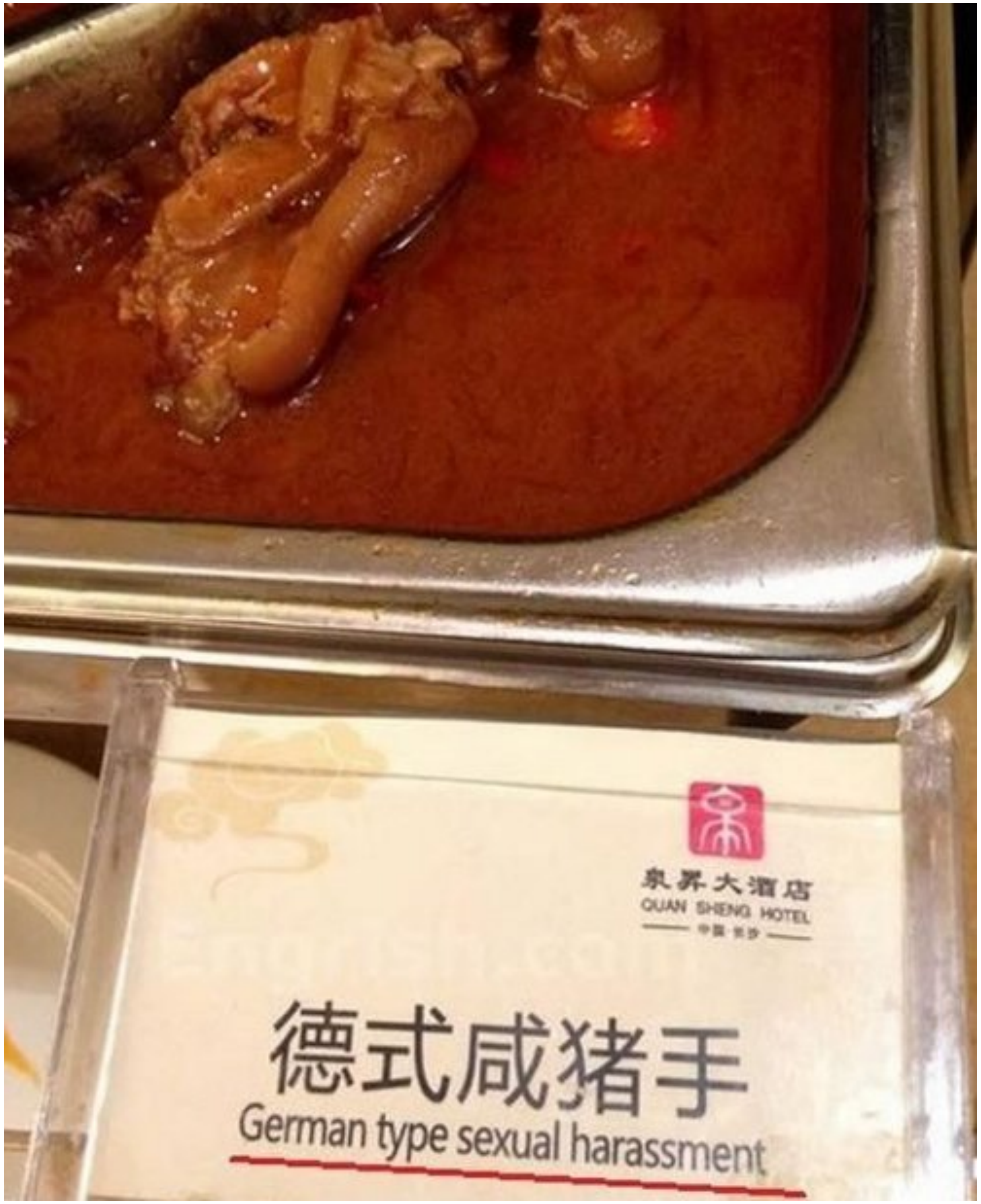
plied by North Country Cooked Meats in Salford Greater Manchester. Th Good Food Chain has si ceased production.

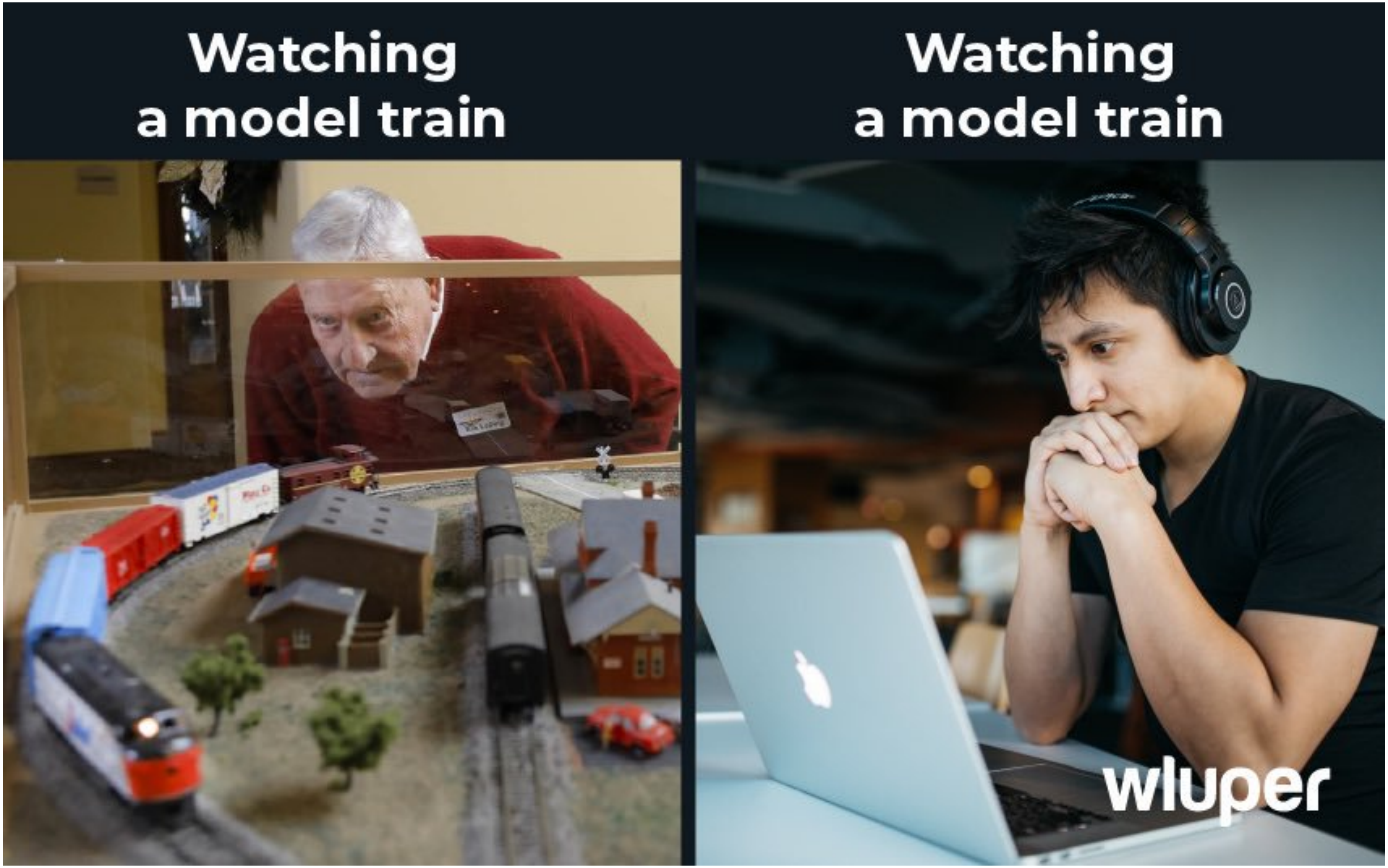
Last night it em that the Food St Agency endorsed given to hospita

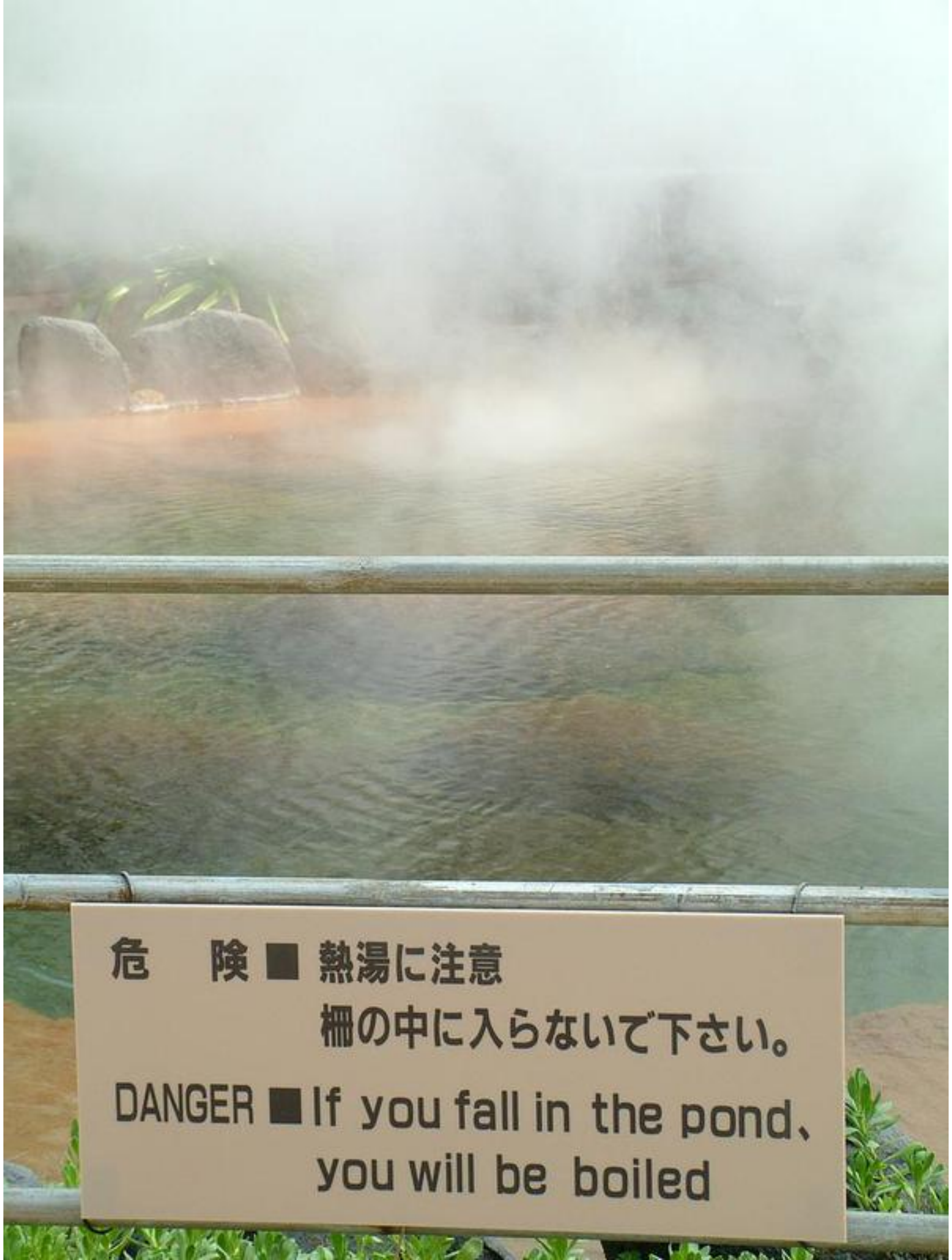
10:56 nachm. · 12. Okt. 2019 · Twitter for iPhone

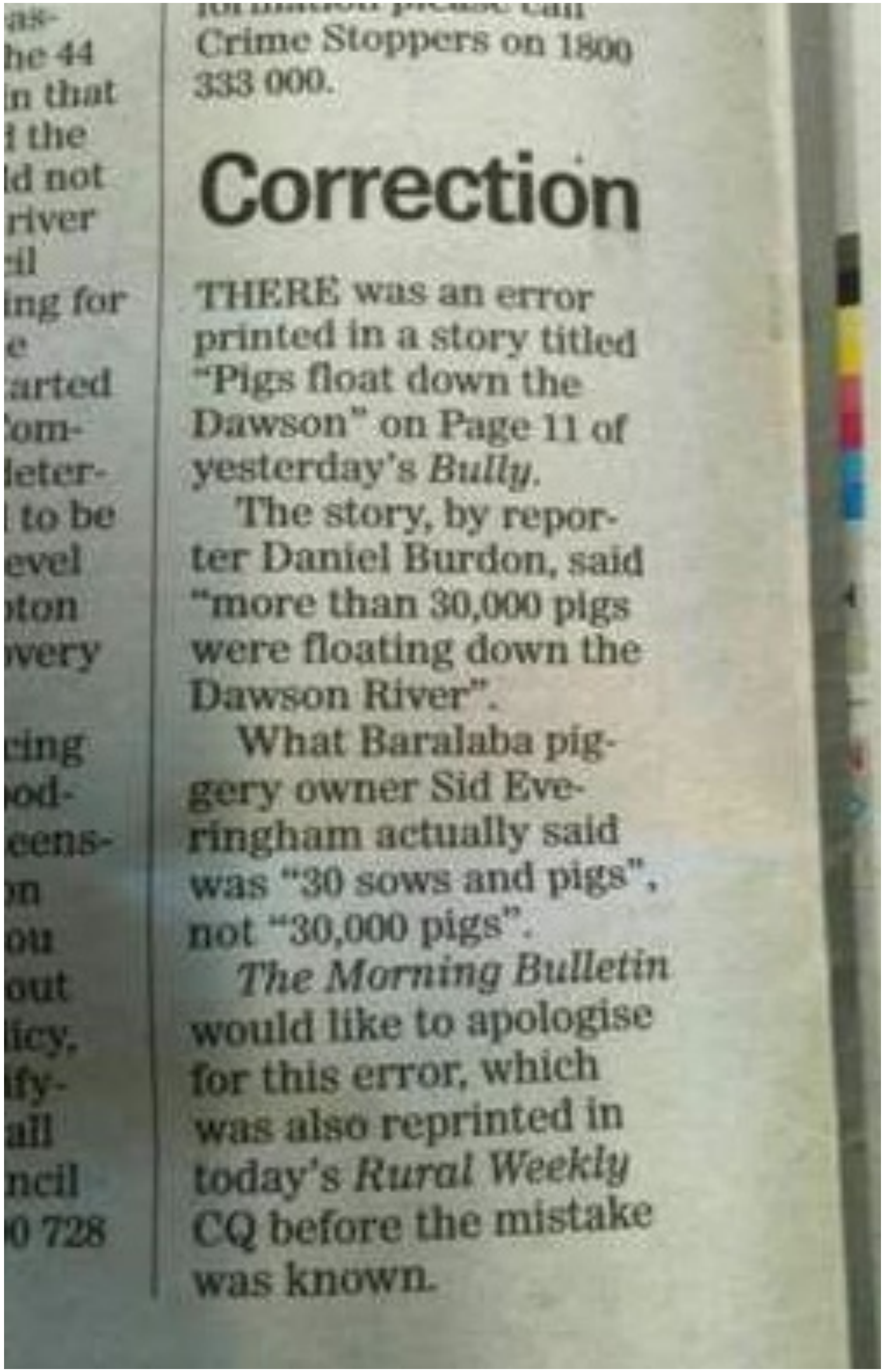


Shared by
Grammarly


泉昇大酒店
QUAN SHENG HOTEL
中国·长沙
德式咸猪手
German type sexual harassment


Watching a model train
Watching a model train
wluper


危　険 ■ 熱湯に注意
柵の中に入らないで下さい。
DANGER ■ If you fall in the pond、
you will be boiled



as-
he 44
n that
the
d not
river
il
ng for
e
arted
om-
leter-
to be
evel
ton
very

cing
od-
eens-
n
ou
out
icy,
fy-
all
ncil
0 728

formation please call Crime Stoppers on 1800 333 000.

Correction

THERE was an error printed in a story titled "Pigs float down the Dawson" on Page 11 of yesterday's *Bully*.

The story, by reporter Daniel Burdon, said "more than 30,000 pigs were floating down the Dawson River".

What Baralaba piggery owner Sid Everingham actually said was "30 sows and pigs", not "30,000 pigs".

The Morning Bulletin would like to apologise for this error, which was also reprinted in today's *Rural Weekly CQ* before the mistake was known.

Dogs must be carried

Hard hats must be worn

why is this hard?

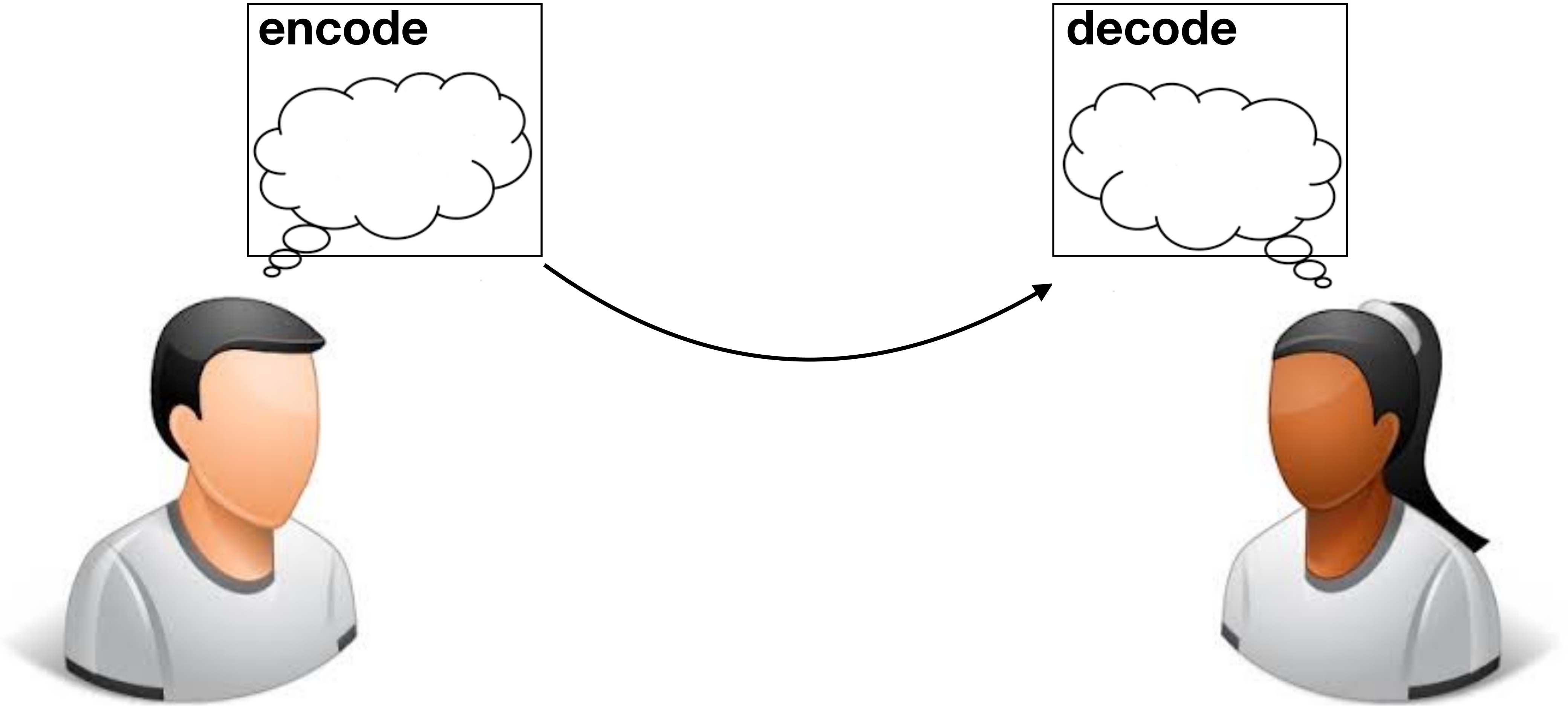
why not just write a simple program that realises NL → R?

- natural language is massively ambiguous… and often it takes the context to identify the intended meaning

- natural language is productive
  “On a sunny May afternoon, Winfried ate a peanut butter sandwich with tuna, while his dog happily sang a song.”

  - you’ve never heard that (I think??), but still you can understand it

a word of caution



Framing encourages a view of language use that abstracts away from language users, and the concrete situations they are in *and make use of*.

NLP and…

- … Computational Linguistics: NLP is the applied cousin? CL has language as object, NLP has language as obstacle to overcome…

- … Machine Learning: Modern NLP uses ML methods. But language does (should? might?) add biases. ML uses language data to test learning algorithms. But language data is peculiar: discrete, skewed distribution (power law), recursive structure.

- … Artificial Intelligence: Used to mean a lot these days (IF-THEN-ELSE?).. Has applied side, to which NLP belongs. Originally, had "intelligence" as object. Brings in common sense reasoning, but also embodiment, spatial reasoning, etc. Turing test is language use test.

NLP and…

- … Speech Processing: Primary form of language is spoken language (and interaction). Often done in different departments, but is coming together more.
- … Humanities: NLP methods in Digital Humanities; Sociology (social media), etc etc.
- … Society: Results have become (superficially) impressive, can be used for good and bad. Be aware.

some central concepts of this class

- knowledge / learning
- representation
- search

knowledge

- to understand language, you must have certain knowledge
- what is “language knowledge”, and what is “world knowledge”?
- how do we get this knowledge into the computer?
    - write it down first, and then formalise it
    - let computer *induce* the knowledge by exposing it to examples of its use

representation

- maths: a collection of objects Y represents a collection of objects X, if it preserves relations between y_i s.t. they conform in some way to relations between x_i
- representation *systems*. Can be more or less *useful.*
  - Example (Marr 1982): Roman numerals vs Indian numerals (decimal system, or binary system, …)
- how shall we represent NL?

How do we represent language?

Text

Labels

the movie was good +

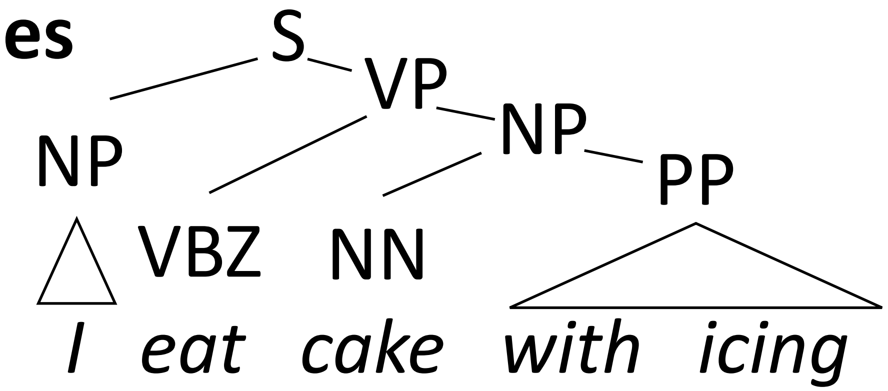
Beyoncé had one of the best videos of all time **subjective**

Sequences/tags

PERSON **WORK_OF_ART**

Tom Cruise stars in the new Mission Impossible film

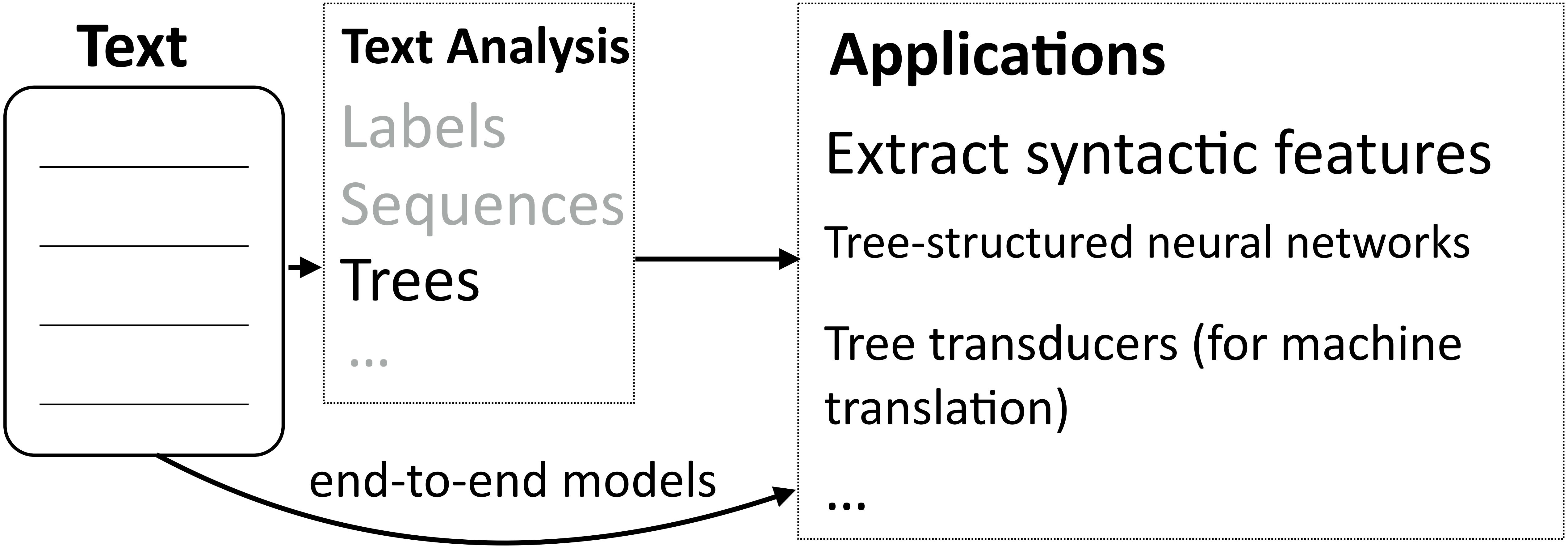
Trees



$\lambda x.\ flight(x) \wedge dest(x)=Miami$

flights to Miami

How do we use these



Text

Text Analysis

Labels

Sequences

Trees

…

Applications

Extract syntactic features

Tree-structured neural networks

Tree transducers (for machine translation)

…

end-to-end models

- Main question: What representations do we need for language? What do we want to know about it?
- Boils down to: what ambiguities do we need to resolve?

search

- how do we efficiently find the object of interest? (often, argmax)

- Viterbi / dynamic programming, CKY, …

some central concepts of this class

- knowledge / learning
- representation
- search

administrativia / this class

- serious time commitment! (9 ECTS = 270 hours of work this semester; of which 45 are contact hours)
- lectures
- 6 assignments
- final (collaborative) project

administrativia / this class

- register on PULS both for Lecture and Lab (“Übung”).
- we won’t make a strict distinction between both
- register on Moodle: `ANLP1991PLNA`
  - forum for discussions
- TAs:
  - Patrick Kahardipraja
  - Sören Etler
- website: https://compling-potsdam.github.io/wise19-bm1-anlp/

administrativia / this class

- lectures on Wednesdays and Thursdays, mixed (see schedule on website)
- time for discussions during lectures as well
- for each assignment, there will be one dedicated lab session to discuss solutions

administrativia / this class

I will mostly follow these textbooks:

- Dan Jurafsky & James H. Martin, Speech and Language Processing, 3rd edition (draft available online. Rerences to this are written as `JM.3.i.k` in the Schedule, with `i` being the chapter.)

- Jacob Eisenstein, Introduction to Natural Language Processing, MIT Press (draft available online). Referred to as `E.chapter`.

Also useful as background reading:

- Yoav Goldberg, Neural Network Methods for Natural Language Processing, Morgan & Claypoole, 2017 (In the library; will try to get them to buy e-book version.)

administrativia / this class

- six programming assignments
- at least 5 must be turned in
- the best 2 from first half & the best 2 from second half are graded
- the sum of the 4 assignments must be at least 250 (of a possible 400) to pass the course
- start early! late assignments will not be accepted

administrativia / this class

final project:

- similar in scope to two assignments, to be completed during break
- group project (2-3 people per group)
- propose and present a topic in the second half of the course
- submission: documented code + two individual papers:
    - planning paper (individual)
    - project presentation (group)
    - implemented project (group)
    - project report (individual)

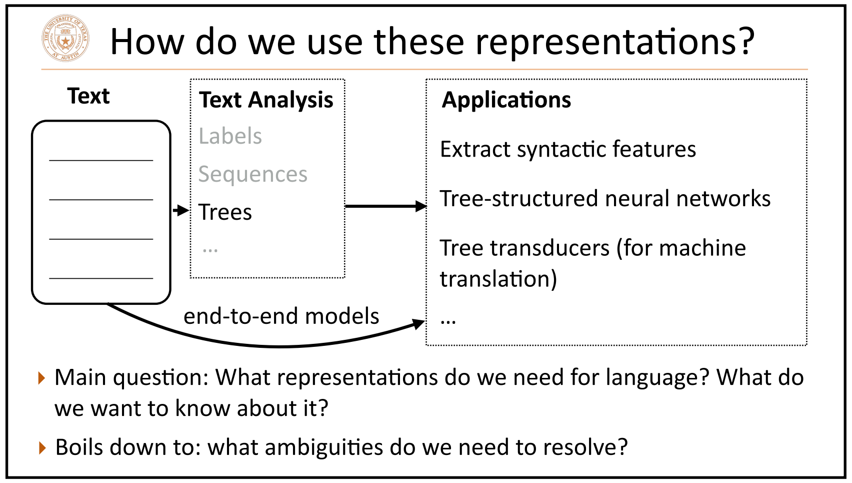
administrativia / this class

- schedule: https://compling-potsdam.github.io/wise19-bm1-anlp/schedule/
- tomorrow: review of probability theory. (Eisenstein, Appendix A.)

Questions, Queries, Comments?

slide credits

slides that look like this



How do we use these representations?

Text

Text Analysis
Labels
Sequences
Trees
...

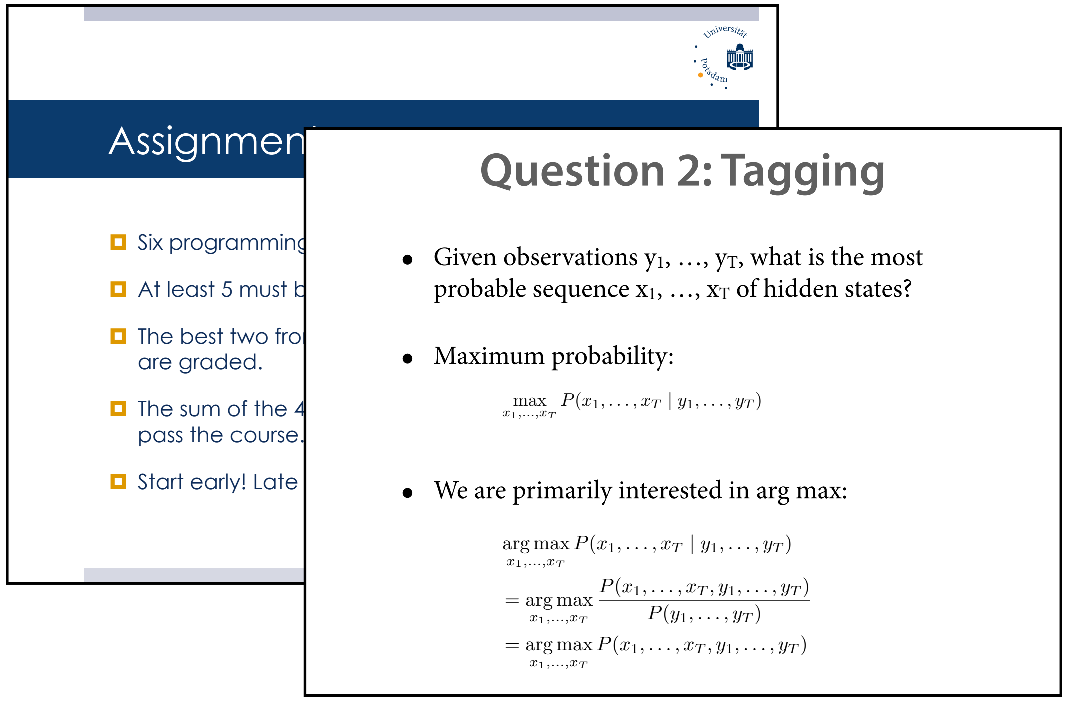
Applications
Extract syntactic features
Tree-structured neural networks
Tree transducers (for machine translation)
...

end-to-end models

- Main question: What representations do we need for language? What do we want to know about it?
- Boils down to: what ambiguities do we need to resolve?

come from

CS388 given by Greg Durrett at U Texas, Austin



Assignmen

- Six programming
- At least 5 must b
- The best two fro are graded.
- The sum of the 4 pass the course.
- Start early! Late

Question 2: Tagging

- Given observations y_1, …, y_T, what is the most probable sequence x_1, …, x_T of hidden states?
- Maximum probability:

$$\max_{x_1,\ldots,x_T} P(x_1,\ldots,x_T \mid y_1,\ldots,y_T)$$

- We are primarily interested in arg max:

$$\begin{aligned}&\arg\max_{x_1,\ldots,x_T} P(x_1,\ldots,x_T \mid y_1,\ldots,y_T)\\ &= \arg\max_{x_1,\ldots,x_T} \frac{P(x_1,\ldots,x_T,y_1,\ldots,y_T)}{P(y_1,\ldots,y_T)}\\ &= \arg\max_{x_1,\ldots,x_T} P(x_1,\ldots,x_T,y_1,\ldots,y_T)\end{aligned}$$

earlier editions of this class (ANLP), given by Tatjana Scheffler and Alexander Koller

and their use is gratefully acknowledged. I try to make any modifications obvious, but if there are errors on a slide, assume that I added them.