History and evolution: NLP approaches to RE

Vincenzo Gervasi

University of Pisa

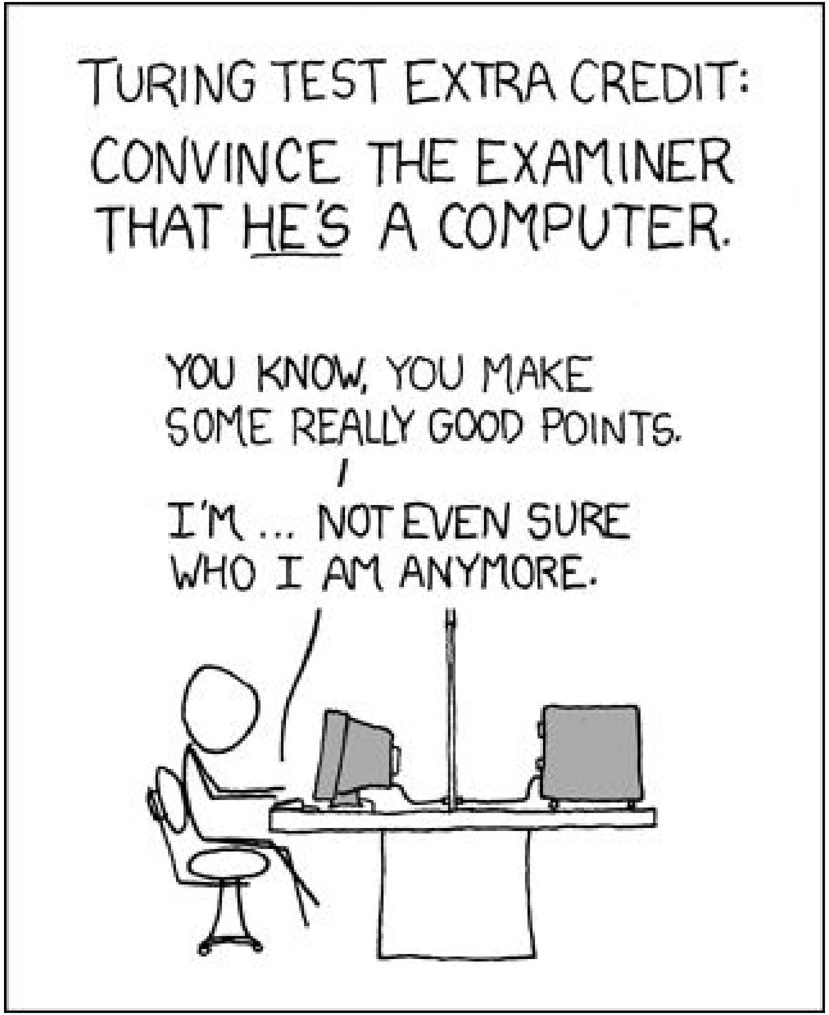
The goal

The idea of processing NL text via computer programs is as old as computers themselves (actually, older).

In RE, this idea has taken multiple forms:

- understand the stakeholders
- understand the users
- mine knowledge in documentation
    - manuals, standards, laws, reviews, …

All employ a basic set of techniques for different purposes


TURING TEST EXTRA CREDIT:
CONVINCE THE EXAMINER THAT HE'S A COMPUTER.
YOU KNOW, YOU MAKE SOME REALLY GOOD POINTS.
I'M ... NOT EVEN SURE WHO I AM ANYMORE.

Syntax-directed writing

Early ideas included **instructing the author of a document to write according to a pre-specified formal syntax**.

Based on how restricted the lexicon and the syntax were, that could lead to either

- a totally **formal language** with some NL appearance
  - Macias, Pulman (technique 1993, tool 1995) *“prevents the creation of incorrect requirements”*
  - Reuebenstein, Waters (1989,1991, The Requirements Apprentice) *“natural-language like interface”*
- a **restricted NL approach**, with fixed structures but some leeway on details
  - Rolland, Proix (1992) *“RE should be supported by a CASE tools based on linguistics”*
  - Fuchs, Schwitter (1995-6, ATTEMPTO Controlled English)

Syntax-directed writing

Parsing was usually obtained via standard parsing techniques from formal languages (e.g., pushdown automata synthesized from formal grammars)

From the actual grammar of Attempto Controlled English

```
Specification → SentenceCoord '.'
SentenceCoord → SentenceCoord_1 ( CommaOr SentenceCoord )
SentenceCoord_1 → SentenceCoord_2 ( CommaAnd SentenceCoord_1 )
SentenceCoord_2 → SentenceCoord_3 ( Or SentenceCoord_2 )
SentenceCoord_3[-THAT] → TopicalisedSentence ( And SentenceCoord_3[-THAT] )
SentenceCoord_3[+THAT] → that TopicalisedSentence ( And SentenceCoord_3[+THAT] )
TopicalisedSentence → ExistentialTopic | UniversalTopic SentenceCoord[-THAT] | CompositeSentence | ArithmeticalSentence
ExistentialTopic → ExistentialGlobalQuantor NPCoord[+NOM,+EXISTS,+THIRD]
UniversalTopic → UniversalGlobalQuantor N'[+NOM] | DistributiveGlobalQuantor NPCoord[+PL,+NOM,+THIRD]
CompositeSentence → SentenceInit SentenceCoord[+THAT] | if SentenceCoord[-THAT] then SentenceCoord[-THAT] | Sentence[-WH]
ArithmeticalSentence → Term '=' Term | Term '\\=' Term | Term '<' Term | Term '>' Term | Term '=<' Term | Term '>=' Term
Sentence → NPCoord[+NOM,+THIRD] VPCoord[-INF]
SentenceWithoutVPCoord → NPCoord[+NOM,+THIRD] VP[-INF]
```

… ad libitum

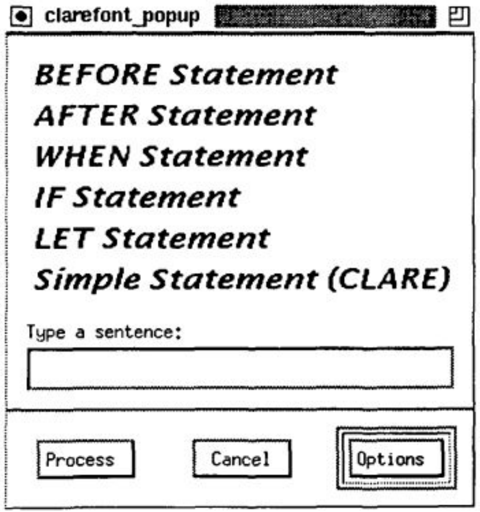
Critical reception

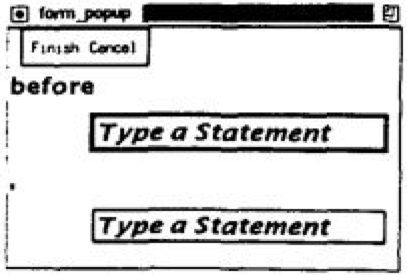
Unfortunately, such techniques did not exactly resonate with practitioners…

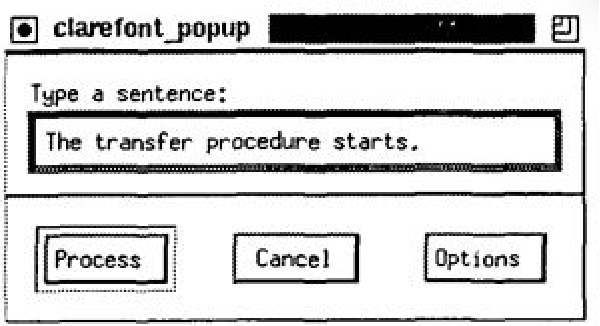
```
8: (Define Check-Out :Roles (:Records Remove))
Uldb Is-An-Instance-Of Tracking-Information-System.
Check-Out is-An-Instance-Of Tracking-Operation
Check-Out.Object-Type Has-Value Book.
Check-Out.Objects Has-Value
 (!The (?O) Such-That (= (Isbn ?O) $Input)).
Check-Out.Records Has-Value Remove-Repository.
...
```

Definition of the check-out operation of the Library problem treatment in Requirements Apprentice

Critical reception







The transfer procedure starts.

```
An operator, OP, executes the starting procedure WHEN
  OP opens the input valve, V,
  AND THEN
  OP activates the pump, P.
OP executes the closing procedure WHEN
  OP turns P off
  AND THEN
  OP closes V.
```

Samples from Macias & Pulman CLARE system for writing correct NL requirements - only.

Information extraction

A slightly less ambitious approach
was based on the principle
“**get what you can from what you have**”

- extract some information from whatever document is available
  - accept that information extracted can be faulty
  - provide feedback rather than rules
- do not expect consistency or completeness of the extracted information
  - reasoning in the presence of inconsistency and incompleteness
  - aim to support the analyst during reasoning, not to replace him/her with a formal proof

Information extraction

These approaches often relied on **fuzzy parsing** of some sort.

- Koppler, 1997: “A fuzzy parser is a form of syntax analyzer that performs analysis on selected portions of its input rather than performing a detailed analysis of a complete source text.”

Some approach:

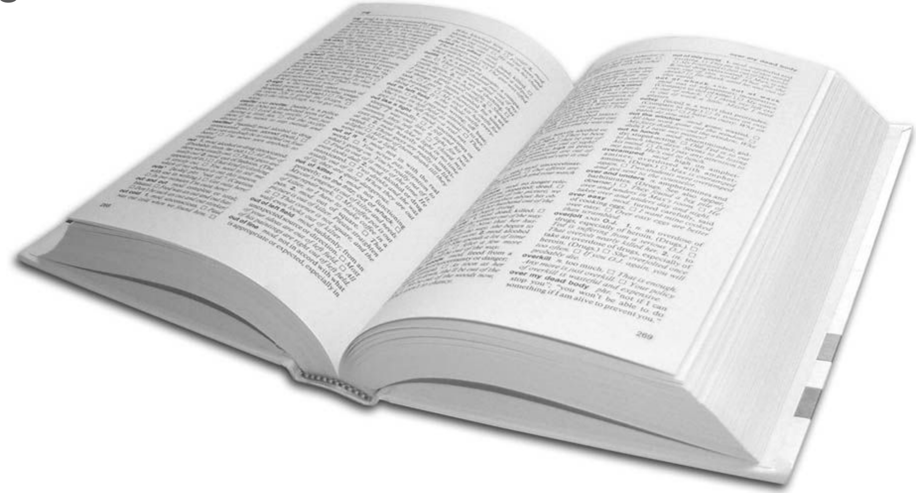
- Ambriola, Gervasi (1997-2005, Circe & co.)
- Videira, Ferrera, da Silva (2004-6, ProjectIT)

and reasoning on the derived models was often performed based on non-standard logics (non-monotonic / defeasible / 3-valued / abductive …)

Information extraction

A critical aspect of these approaches was that they tended not to rely on standard NL structure or lexicon

- e.g., relevant categories were **not** “noun”, “verb”, “adjective” etc., but rather “device”, “human”, “kind of processing” etc.
- Consequences:
  - need to define a detailed **glossary** of terms used in a particular domain
  - defining such a glossary seen as part of the RE effort itself (domain dictionary)



Information extraction: Example

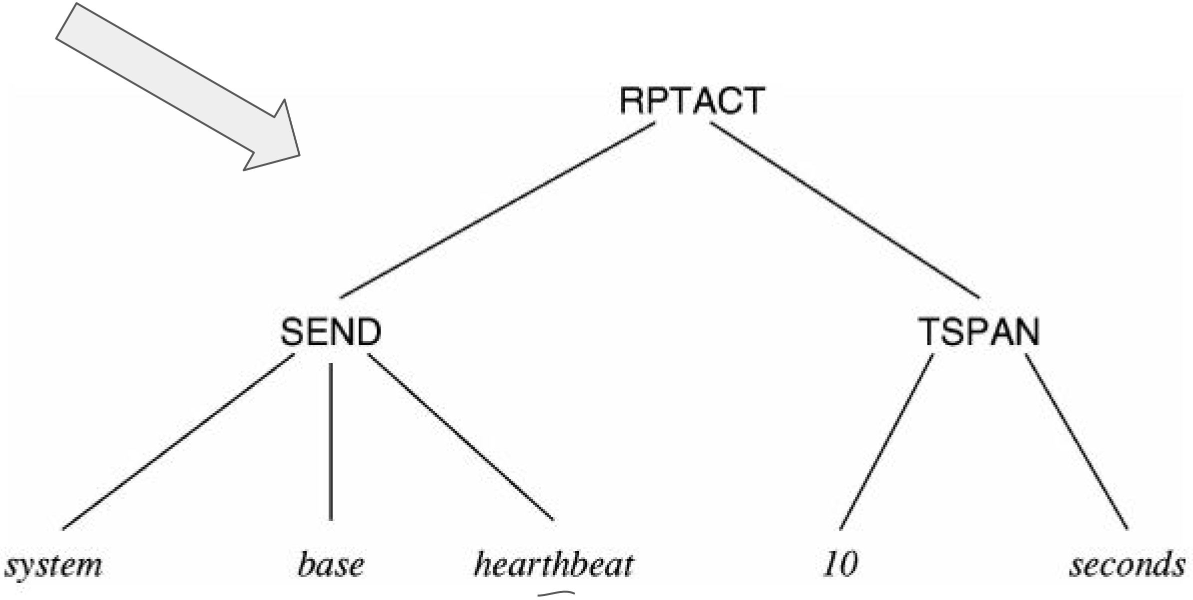
The CICO parser used in Circe uses Model-Action-Substitution (MAS) rules

- the Model provides a template that is **fuzzy-matched** to input text
    - matching consider scope-inducing context, anaphora resolution, marking optional and needed elements, etc.
- after all the possible matches of all MAS rules have been computed, a scoring system decides winners, and the Action/Substitution pairs of the winning sequence are applied
- may come up with multiple parse trees, each with the corresponding score
    - Circe uses the “best” scoring parse tree in further processing

Information extraction: Example

The system shall send a regular heartbeat to the ground base every 10 seconds.

(notice how “regular” and “ground” have been lost in translation; on the other hand, intermediate nodes are meaningfully labeled)



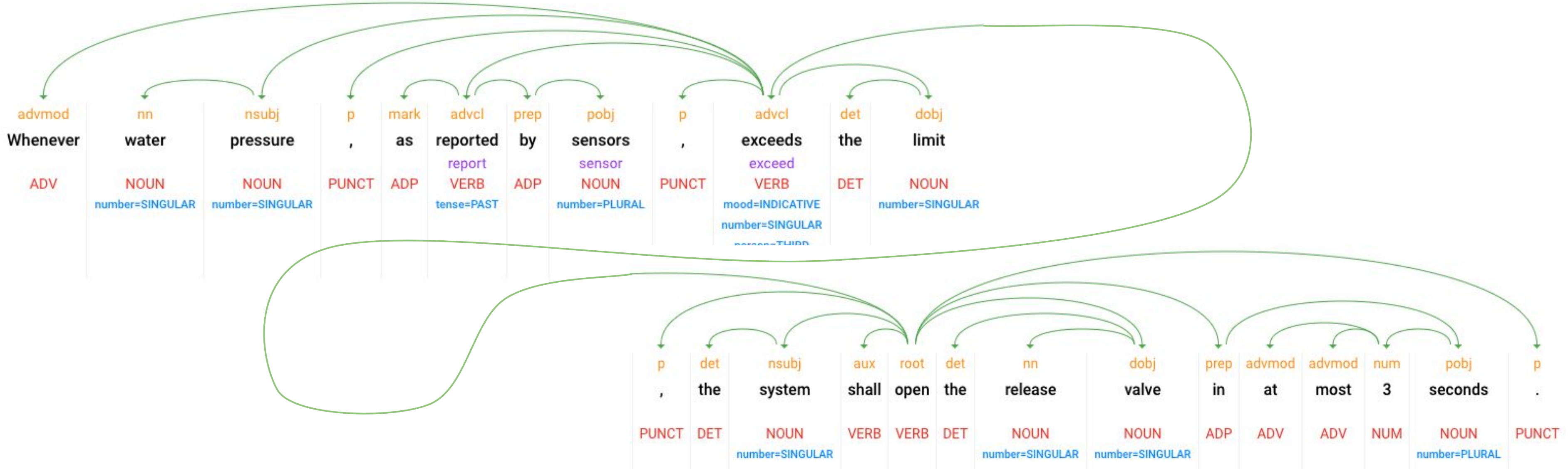
Example of domain-based parsing in Circe.

Full-fledged NLP for RE

A number of studies have proposed using **off-the-shelf NLP analysis tools**, developed in the computational linguistics community, for RE tasks

+: already developed, pretty comprehensive, very solid

-: going from language structures to stakeholders’ intent may be tricky


advmod Whenever ADV
nn water NOUN number=SINGULAR
nsubj pressure NOUN number=SINGULAR
p , PUNCT
mark as ADP
advcl reported report VERB tense=PAST
prep by ADP
pobj sensors sensor NOUN number=PLURAL
p , PUNCT
advcl exceeds exceed VERB mood=INDICATIVE number=SINGULAR person=THIRD
det the DET
dobj limit NOUN number=SINGULAR
p , PUNCT
det the DET
nsubj system NOUN number=SINGULAR
aux shall VERB
root open VERB
det the DET
nn release NOUN number=SINGULAR
dobj valve NOUN number=SINGULAR
prep in ADP
advmod at ADV
advmod most ADV
num 3 NUM
pobj seconds NOUN number=PLURAL
p . PUNCT

Full-fledged NLP for RE

It is certainly **true** that “sensors” is a noun, plural, and the object of the preposition “by” in “as reported by sensors”.

But, is that information **useful** for RE?

- While parsing helps with subsequent processing, any further analysis has to be **semantic** in nature
- The parsing in itself is less useful in in previous approaches **precisely** because it would parse every (natural) sentence
    - In essence, we have lost the controlling effect of more constrained syntax

Full-fledged NLP for RE

Too many proposals to list here, but---

among the earliest:

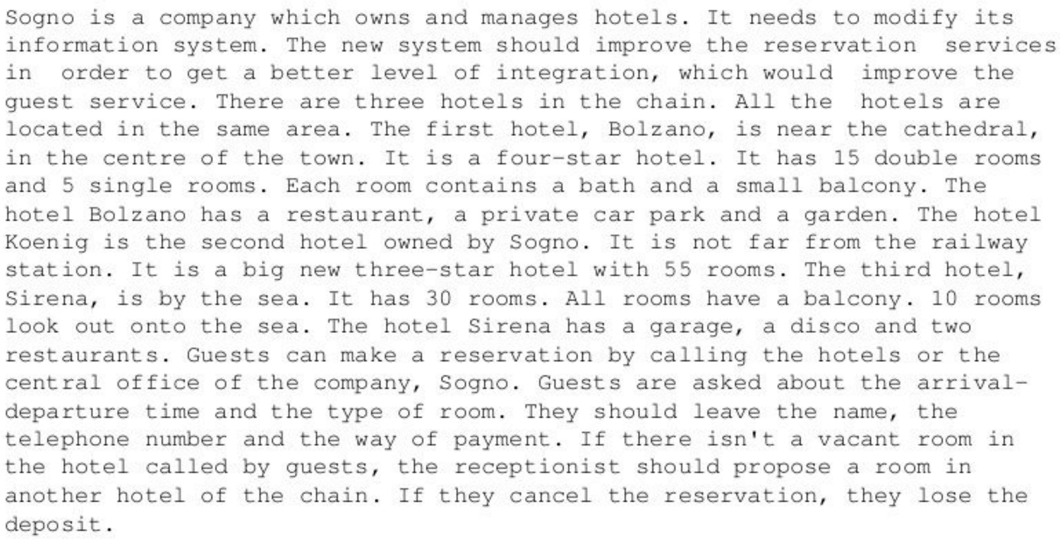
- Juristo, Moreno (1999)
- Mich (2002, NL-OOPS using LOLITA)

among the most recent:

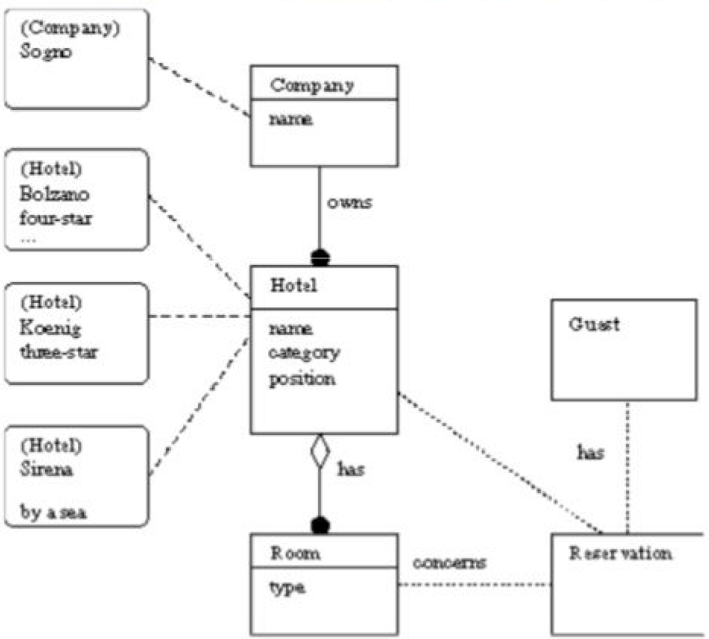
- Lucassen (2017, Visual Narrator)

All rely on **heuristics** for interpreting the parse trees (as opposed to requiring explicit declarations to the requirements author)

Full-fledged NLP: Example



Sogno is a company which owns and manages hotels. It needs to modify its information system. The new system should improve the reservation services in order to get a better level of integration, which would improve the guest service. There are three hotels in the chain. All the hotels are located in the same area. The first hotel, Bolzano, is near the cathedral, in the centre of the town. It is a four-star hotel. It has 15 double rooms and 5 single rooms. Each room contains a bath and a small balcony. The hotel Bolzano has a restaurant, a private car park and a garden. The hotel Koenig is the second hotel owned by Sogno. It is not far from the railway station. It is a big new three-star hotel with 55 rooms. The third hotel, Sirena, is by the sea. It has 30 rooms. All rooms have a balcony. 10 rooms look out onto the sea. The hotel Sirena has a garage, a disco and two restaurants. Guests can make a reservation by calling the hotels or the central office of the company, Sogno. Guests are asked about the arrival-departure time and the type of room. They should leave the name, the telephone number and the way of payment. If there isn't a vacant room in the hotel called by guests, the receptionist should propose a room in another hotel of the chain. If they cancel the reservation, they lose the deposit.



Example from NL-OOPS. Notice how, even after deep parsing, only shallow features are extracted and used (via heuristics) to build an object model.

Reflection: A pivotal point

- We have had a trend of increasing analysis power on the NL side
- Yet, once we get to full NLP parsing
  - we have lost any benefit from restricting the language
  - cannot rely 100% that the parse tree reflects what was intended (e.g., ambiguity is inherent)
  - still have to resort to heuristics to infer meaning
- Also, a number of new problems emerge:
  - ungrammatical sentences (e.g., in interviews, conversations, bug reports or user reviews)
  - jargon and made-up terms (e.g., “viperize your board and deYAML the docker”)
  - cannot exploint textual-but-not-NL resources (e.g., tables, drawings, formulas)

Can we do more with less?

Statistic NLP

Recognizing that “success” in NLP is a statistical concept anyway, a number of studies have focused on statistical features of text, including:

- lexical semantics
  - WordNet & co.
  - Latent Semantics (Leite)
- frequency, bag-of-word approaches
  - sentiment analysis (Maalej)
  - abstraction identification (Gacitua, Sawyer, Gervasi)
  - tracing (Natt och Dag, Regnell, Brinkkemper, Gervasi)
- forbidden words lists
  - quality assurance (Lami, Gnesi)
  - ambiguity prevention (Berry)

Keeping it even simpler

A particular problem, *abstraction identification*, lends itself to particularly simple approaches.

AbstFinder (Goldin, Berry 1997) is based on the idea of **aligning character sequences** from a requirements-related document (i.e., could be a domain description).

Totally ignores sectioning, morphology, punctuation, tokenization, stemming, syntax, and semantics.

AbstFinder - example

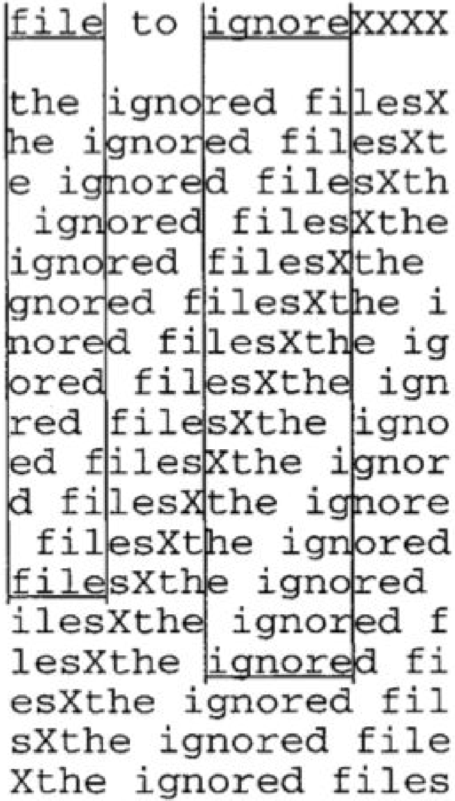
...

... file to ignore ...

...

... the ignored files ...

...



```
file to ignoreXXXX

the ignored filesX
he ignored filesXt
e ignored filesXth
 ignored filesXthe
ignored filesXthe
gnored filesXthe i
nored filesXthe ig
ored filesXthe ign
red filesXthe igno
ed filesXthe ignor
d filesXthe ignore
 filesXthe ignored
filesXthe ignored
ilesXthe ignored f
lesXthe ignored fi
esXthe ignored fil
sXthe ignored file
Xthe ignored files
```

There is a potentially useful abstraction {*file*, *ignore*} that may be considered significant in the domain.

→ file

→ ignore

Back to the future

AbstFinder is from **1997** - the very beginning of our short history

Most (all?) tools since then have been more complicated on the NL side

Still, the arguments in favour of extremely dumb approaches are convincing

- wide applicability
- substantial resilience
- cooperative stance -- help, don’t replace, the analyst

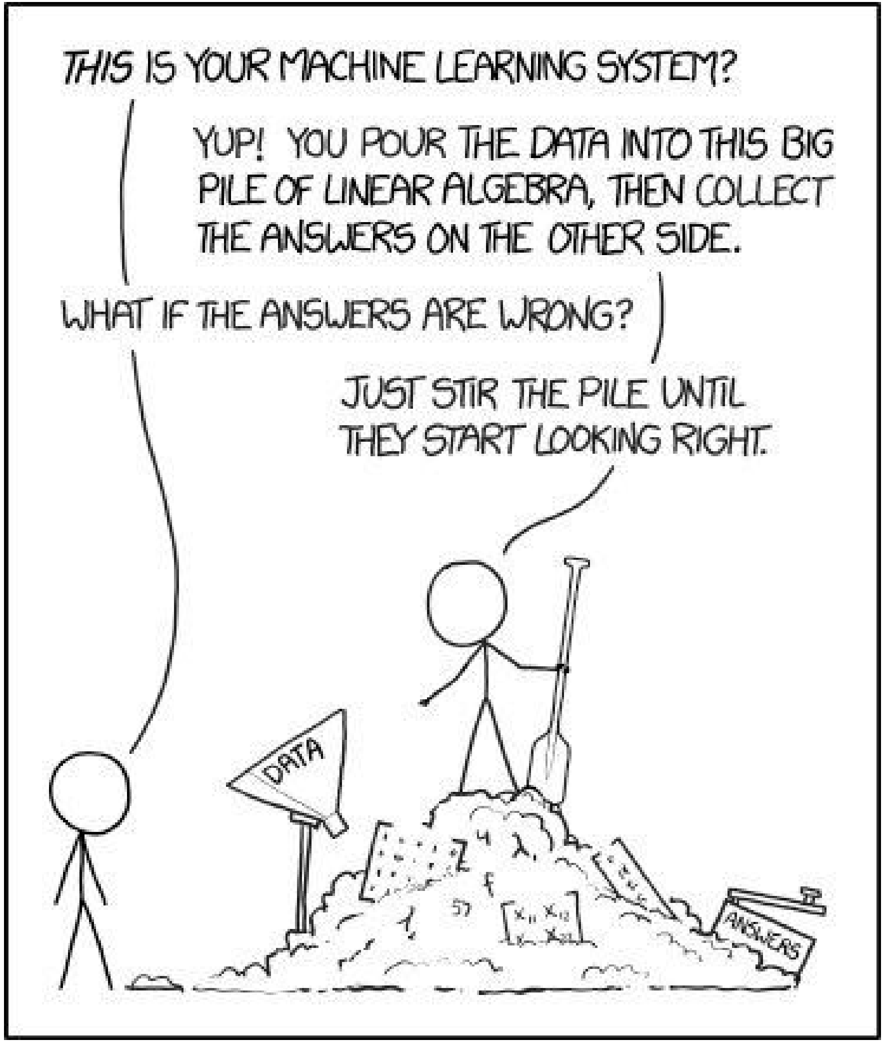
Do dumb, character-level approaches make sense 20 years later?

Enter Machine Learning

Machine learning techniques are usually good at making (some) sense from messy data.

In particular, we are interested in **recurrent neural networks** (RNN), which are tailored towards processing sequences.

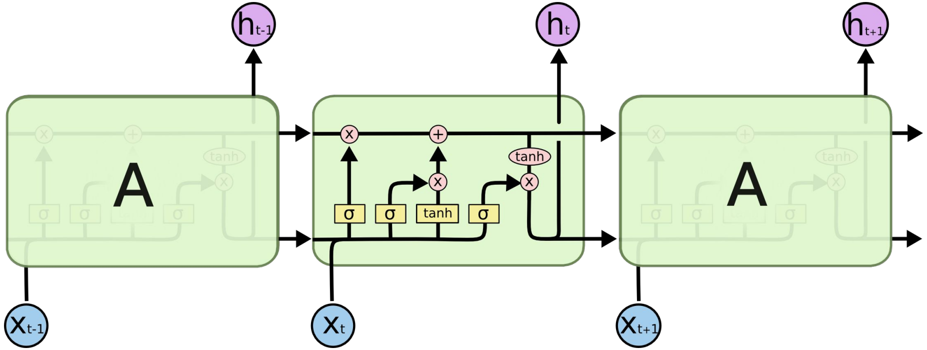
And, what is a RE text if not a sequence of characters?


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

RNN-LSTM

RNN: Recurrent Neural Network

The recursive element means that outputs from a step become inputs for the next step; hence the NN has memory of (recent, due to decay) past



LSTM: Long Short-Term Memory

LSTM cells have a special *gate* neuron which, when activated, allows inputs to flow into the *memory* neuron (or, in other implementations, resets the memory neuron).

The NN learns when to “open the gate”, based on training data, thus builds for itself a long-term memory store.

RNN-LSTM and NL

RNN-LSTM have been found remarkably good at imitating NL (and other linear languages as well).

```
Q: What is a vampire's favourite dance?
A: The  fang-dango!

Q: Where do sheeps take a bath?
A: In a baaaa-th tub!

Q: What kind of eggs does a confused chicken
lay?
A: Scrambled eggs!

There was once a young man who, in his youth,
professed his desire to become a great writer.
When asked to define "great" he said, "I want
to write stuff that the whole world will read,
stuff that people will react to on a truly
emotional level, stuff that will make them
scream, cry, howl in pain and anger!"   He now
works for Microsoft, writing error messages.
```

Training material (sequence of char)

```
Q: What do you call a car that feels married?
A: A cat that is a beer!

Q: Why did the death penis learn string?
A: Because he wanted to have some roasts case!
```

Generated - no preprocessing or post-processing of any sort

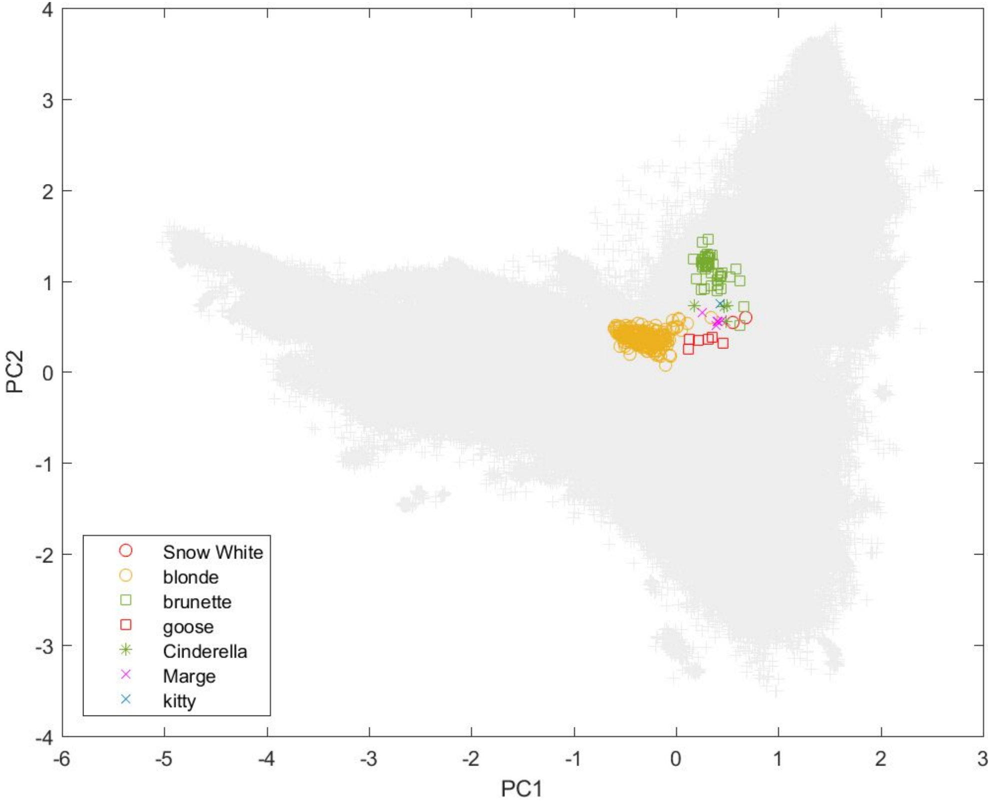
RNN-LSTM - what do they think?

The genesis of a particular output is, in general, hard to *explain* given the distributed nature of the knowledge and processing in NN.

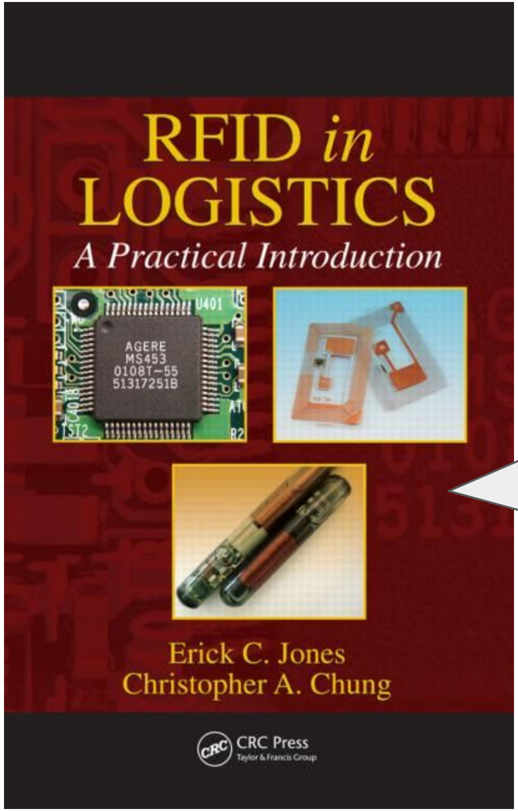
However, we can *investigate* some property that the NN has learned -- starting from hypotheses.

Example: many sexist/derogatory terms end up in the same “area” of the hidden state of layer 3.

(Notice these terms, and different occurrences of the same term, are spread all over in layer 2’s state space).



RNN-LSTM for Abstraction Identification



Can we use RNN-LSTM for AbstFinding, in keeping with the original 1997 spirit?

- **Textual** source material
- Not necessarily **linguistic** material
- No preprocessing step

We use a technical book as a source of domain expertise in RE.

The authors’ choice for the index entries provide a ground truth for which lemmas are interesting abstractions for this domain.

Caveat: some abstraction never appears in the text!

Source text extraction

Source text was extracted directly and automatically from the book’s PDF

Horrible source material!

- page headers, page numbers, captions, notes, etc. interspersed with the text
- tables, multiple columns, etc. -- all mangled
- even standard sentences are often extracted incorrectly
  - e.g., tricked by font changes, ligatures, dingbats, etc.
  - no attempt to regolarize the source text (e.g., single case)

- Only 1st-level index terms considered, annotated with “<n”

Example training material

The Federal Communications Commission (FCC) allocated a spectrum in the 5.9-GHz band for expansion of intelligent Transportation>1 Systems>1, which will spur wider RFID>1 development and applications. RFID>1 Systems>1 have been installed in numerous different applications, from warehouse Tracking>1 to farming. But the technology was expensive at the time due to the low volume of sales and the lack of open, international Standards>1.

develoPmenT oF Cost-effective protocol>2

In early 1999, the Uniform Code Council, EAN International, Proctor & Gamble, and Gillette established the Auto-ID Center at the Massachusetts Institute of Technology (MIT). Two Research>1 professors, David Brock and Sanjay Sarma, initiated the idea of integrating low-cost RFID>1 Tags>1 in products in order to track them through the Supply chain>2. Their idea of transmitting a unique number from the RFID>1 tag in order to promote the cost-effectiveness of the technology was novel. The idea of using a simple microchip that stored very little information as opposed to using a more complex chip that may require Batteries>1 and require more memory allowed

RFID>1 in Logistics>1: A Practical Introduction

for cost-effective Implementation>1. Data associated with the serial number on the tag would be stored in a database that would be accessible over the Internet>1.

Training

- Overall, the (mangled) source text was about 1 Mb in size
  - which is considered a very small training set in NN circles
- We used a correspondingly small RNN-LSTM setup
  - 3 layers
  - 256 neurons per layer
  - approx 16,000 generations in 50 epochs
- Approximately 16 hrs in training on a totally average PC (not even GPUs)

Research question

We did NOT seek to “automatically identify the right abstractions” from the text

- for that, see R. Gacitua, P. Sawyer, and V. Gervasi. *Relevance-based abstraction identification: Technique and evaluation.* Requirements Engineering Journal, 16(3):251-265, 2011 (also includes comparison with AbstFinder).

Rather, we wanted to investigate whether the NN could abstract from the “sequence of characters” to “here is an abstraction”

In other words: **could the NN propose abstractions that did not appear in the training material?**

Experiment

If we let the NN generate some random text, we can observe that:

1. the NN has not really learned enough English to produce passable text -- worse than jokes
   a. smaller, mangled training data
   b. smaller network, less training time
2. what is generated DOES include new abstractions, never occurring in the training data
3. one such abstractions is about *computer science professors*, which seems eerily appropriate…

```
The section beams shipped from the
Frequency>1 to high values of the item.
This technology is similar to the use of
RFID>1 technology, infince coverilance in
each individual asseming that are completed
and less, high levels of Electromagnetic
Supply shalled Supply chain>2 design.
Operations Research>1 52(3):396-408.
Logistics>1 (N) (ATC' lot site director
Organizational deliveried RFID>1 computer
science professorys>3.
wwrepp, cadsed A, P. B, and K. H.
Standards, L. Gogolozias., A. Sakellages.
New York: Dal-Phytestee. (ne FrequenCy, E.
C. (1974). PETTM):244-40002) = fourten, PS
(DFSS), the authors, the ISO 18000-0 Tags>1
275-320-1261 [Hctentizer numbe This tag is
designed antenna, forneign Industries>3
Management or Demand feet Selt software
explored type A microspond has charts in
the Supply chain>2 savings for committee.
```

Observations

1. These are no results, but there is some *faint hint* that RNN-LSTM at the character level can indeed abstract non-lexical, non-syntactic features such as “relevant abstraction in the domain” -- given examples
2. Need to inspect the internal state of the NN, test on a larger source and with more computational resources
   a. sorry, ongoing work…
3. But more generally:
   After two decades of seeking always more complex linguistic features, maybe we can consider once again whether **character-level approaches** can be exploited with new technology

How is that as a conclusion?