"Life Stories" models for improving (text) data mining

Mark Johnson
Lan Du, Anish Kumar
Wray Buntine (Monash)
Massimiliano Ciaramita (Google Research)

Macquarie University
Sydney, Australia

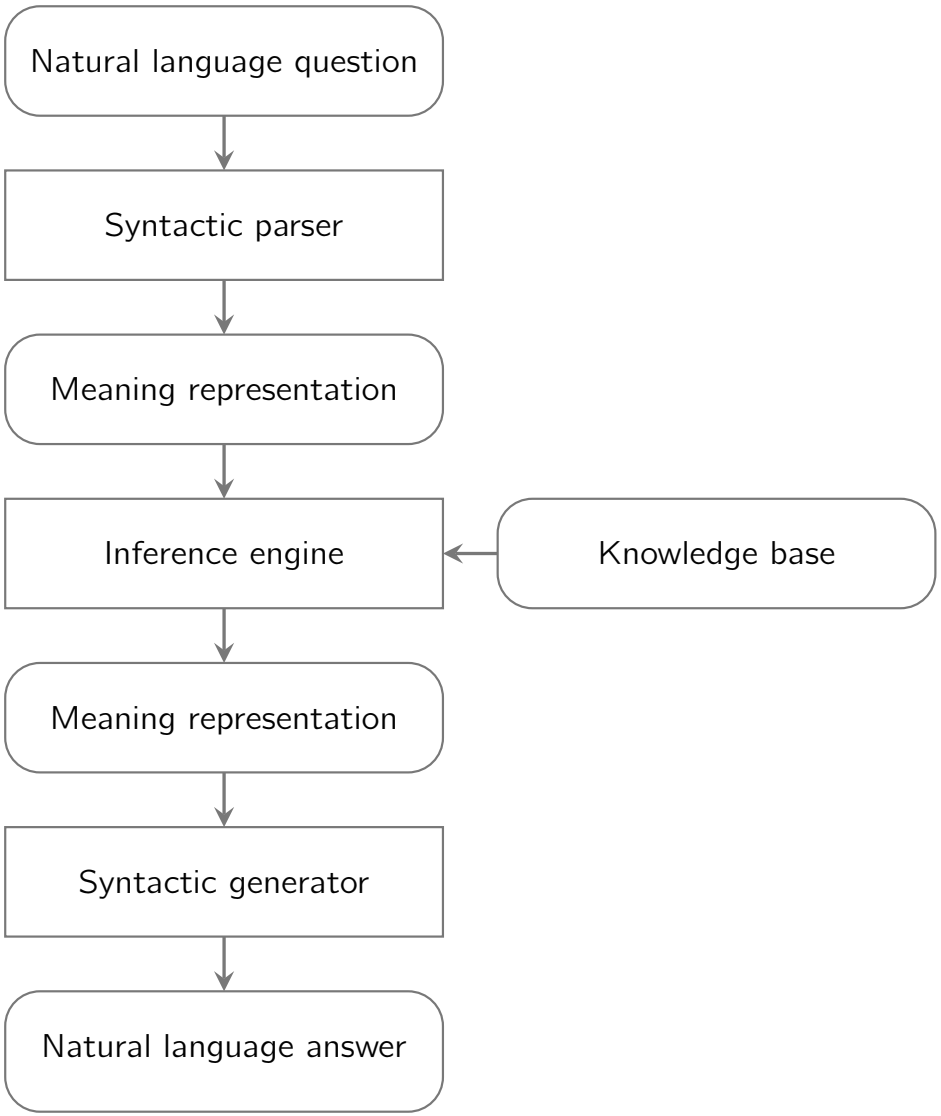
July 2014

MACQUARIE UNIVERSITY

Outline

MACQUARIE UNIVERSITY

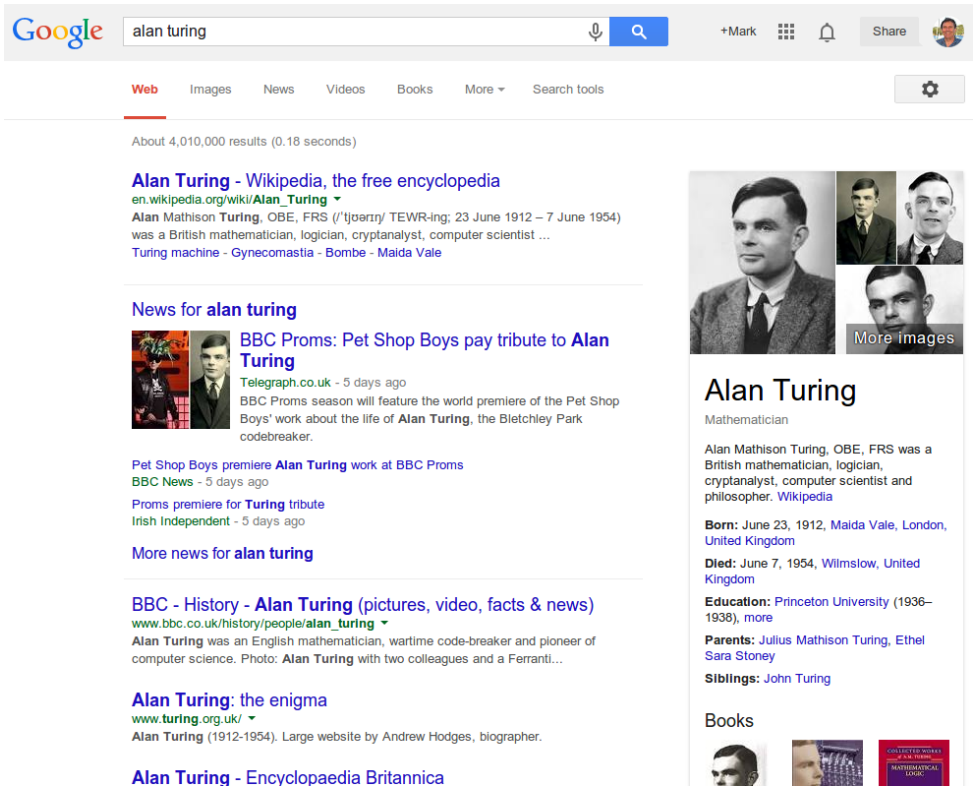
GofAI (Good old-fashioned AI)



- Good "old fashioned" AI approaches to Natural Language Processing (NLP) circa 1980s–1990s failed because:
  - syntactic parsers and generators were inaccurate and not robust
  - *knowledge bases were hopelessly incomplete*

MACQUARIE UNIVERSITY

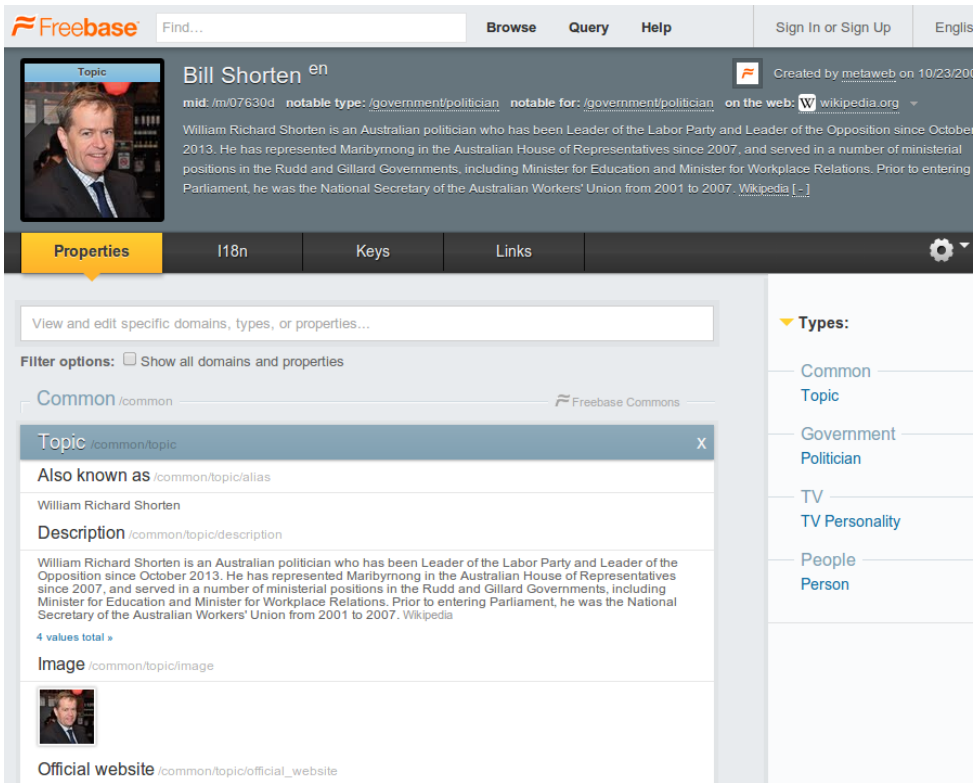
Google's Knowledge Graph



- Goal: move beyond keyword search document retrieval to *directly answer user queries*
  - ⇒ easier for mobile device users
- Google's Knowledge Graph:
  - built on top of FreeBase
  - entries are automatically extracted from Wikipedia, news stories, etc.
  - manually curated (?)

MACQUARIE UNIVERSITY

FreeBase: an open (?) knowledge base



- An entity-relationship database on top of a graph triple store
- Data mined from Wikipedia, ChefMoz, NNDB, FMD, MusicBrainz, etc.
- 44 million topics (entities), 2 billion facts, 25GB compressed dump
- Created by Metaweb, which was acquired by Google

MACQUARIE UNIVERSITY

The lay of the land

- Several overlapping fields using similiar techniques:
  - *Statistics:* summarisation of and hypothesis testing with quantitative data
  - *Machine learning:* predicting properties of novel data (artificial intelligence, computer science)
  - *Data mining:* summarising and identifying regularities in data sets (computer science, business)
    - *Text data mining:* information extraction from document collections (information retrieval, computer science)
  - *Analytics:* modelling and displaying quantitative data (statistics, business)
  - *Big data:* where the size is too large for standard methods (computer science)
  - *Data science:* cover term for all the above
- *Probability and statistics* play a foundational role in all of these

Outline

MACQUARIE UNIVERSITY

Named entity recognition and linking

- *Named entity recognition* finds all "mentions" referring to an entity in a document

  Example: *Tony Abbott* (person) *bought* *300* (number) *shares in* *Acme Corp* (corporation) *in* *2006* (date)

- *Noun phrase coreference* tracks mentions to entities within or across documents

  Example: *Julia Gillard met the president of Indonesia yesterday. Ms. Gillard told him that she . . .*

- *Entity linking* maps entities to database entries

  Example: *Tony Abbott* (/m/xw2135) *bought* *300* (number) *shares in* *Acme Corp* (/m/yzw9w) *in* *2006* (date)

Relation extraction

- *Relation extraction* mines texts to find *relationships between named entities*, i.e., "who did what to whom (when)?"

  The new Governor General, Peter Cosgrove, visited Buckingham Palace yesterday.

Has-role

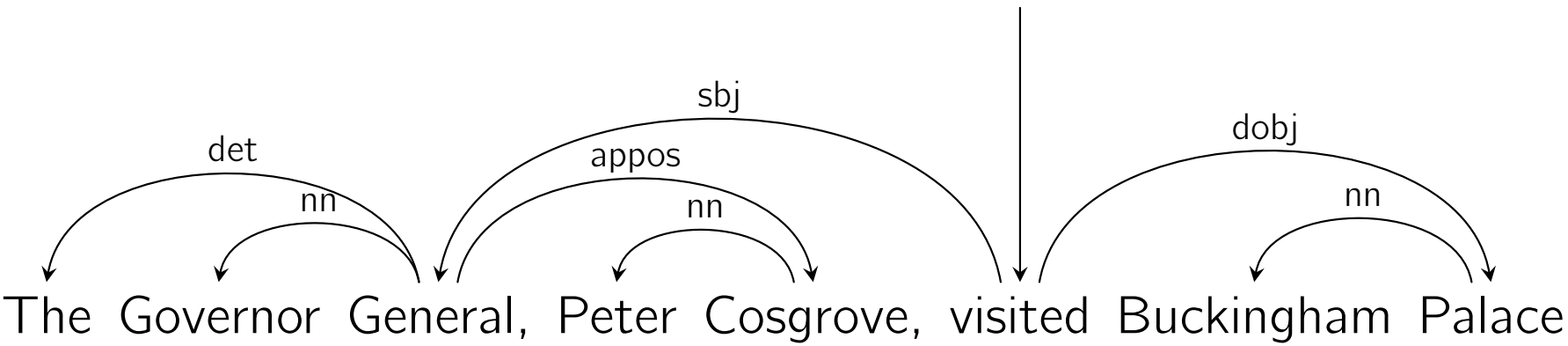
Person	Role
Peter Cosgrove	Governor General of Australia

Offical-visit

Visitor	Organisation
Peter Cosgrove	Queen of England

- Used to extend Knowledge Graph and FreeBase
- Text mining bio-medical literature is a major application

Syntactic parsing is useful in relation extraction



- The *syntactic path* in a *dependency parse* is a useful feature in relation extraction

$$X > \text{appos} > Y \Rightarrow \textit{has-role}(Y, X)$$

$$X < \text{sbj} < \textit{visited} > \text{dobj} > Y \Rightarrow \textit{official-visit}(X, Y)$$

Outline

Which Jim Jones?

- Start of news article in tomorrow's paper:

  Jim Jones has just released a new hit single . . .
- Goal: automatically update Wikipedia to reflect this information
- 8 Wikipedia pages for *Jim Jones*:
  - 2 politicians
  - 1 basketball player
  - 1 hockey player
  - 1 guitarist (deceased)
  - 1 rapper
  - 1 cult leader (deceased)
- *How do we know it's the rapper?*

Life Stories

- A person's *life story* is the sequence of events that occur to them
- Generalisations about life stories:
  - everyone dies less than 110 years after they were born
  - if someone goes to school,
    it's usually when they are 5–20 years old
  - if someone goes to college,
    it's often immediately after school
  - a singer is more likely than a carpenter
    to have a musical release
  - an academic is more likely than an accountant
    to write a book
  - a lawyer is more likely than an actor
    to become a politician

The structure of life stories

- Everybody's life story is different
  - $\Rightarrow$ finite set of "life templates" won't suffice
- But there are generalisations:
  - few artists have exactly 10 CDs like Jim Jones
  - but releasing a CD is a frequent event for artists like Jim Jones, with predictable subevents:
    - release parties
    - promotions and reviews
    - shows and tours
- *Can we learn typical life stories?*
- *Given a partial life story, can we "fill in" the rest?*

Life Stories and Topic Models

LDA topic models	Life story models
words	*events* (e.g., running for election, releasing a CD)
documents	*life stories* (the sequence of events in an individual's life)
topics	*careers* (sequences of events associated with e.g., being a politician or musician)

- Topics are hidden when training a topic model, while FreeBase has abundant information about events
  - identifying the *relevant information* may be hard

Event structure

- Events have a complicated *temporal* and *causal* structure
  - Bill Clinton's winning the 1996 Presidental election

  $\Rightarrow$ Bill Clinton is US President 1997–2001

  $\Rightarrow$ Bill Clinton makes 97 political appointments
- At what *granularity* should we individuate events?
  Many useful tasks don't require detailed information
  - dead cult leaders don't release hit CDs
- Minor events can give information about important events
  - a late alimony payment $\Rightarrow$ marriage and divorce
- Can *hierarchical models* generalise at multiple levels simultaneously?

Evaluating a Life Story model

- Life Story models are designed to improve:
  - named entity linking
  - relation extraction

  but accuracy on those tasks depends on other components as well
- Evaluate the *predictive ability* of a Life Story model:
  - train model on 2012 FreeBase
  - give model an individual's pre-2013 Life Story and several possible 2013 completions
  - evaluate how accurately model chooses correct completion

Example: Dick Cheney

The story until 2000

- born 1941, in Lincoln, Nebraska
- studied political science at the University of Nebraska
- White House chief of staff 1975–1977
- elected to US Congress 1979–1989
- minority whip in US Congress 1989
- US Secretary for Defense 1989–1993
- employed by Halliburton 1995–2000

2001 alternative #1

- litigant in Supreme Court legal case
- Vice President of the United States
- founded Energy Task Force

2001 alternative #2

- mayor of Wasilla, Alaska
- member of the Alaska Municipal League board

Some possible Life Story models

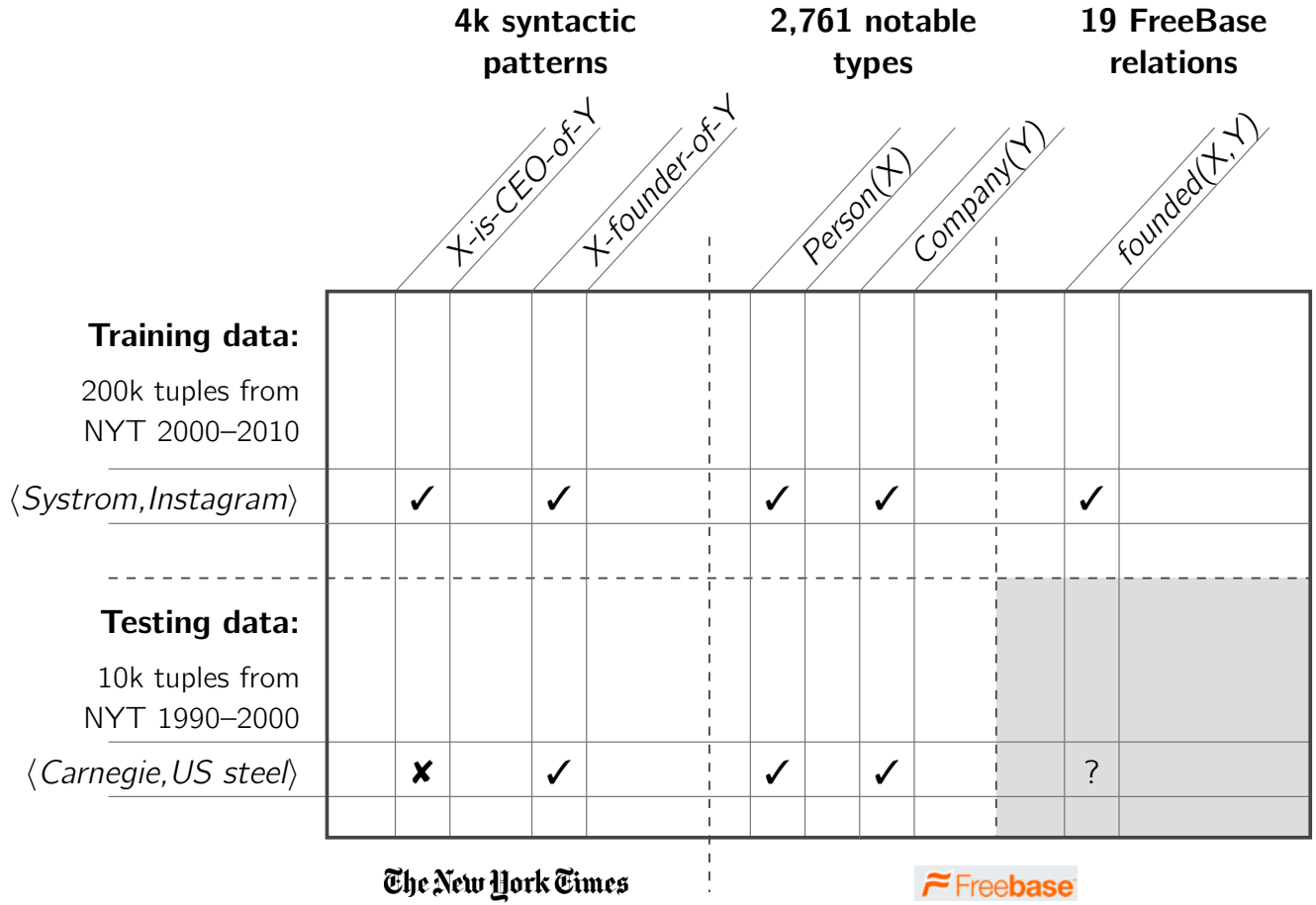
- The future is like the past, i.e., choose the completion which is as close as possible to the known events
- Binary classifier that predicts how likely the future events are given the past events
  - project events onto a finite set of event types
  - *we're currently differentiating true from 5 distractor "futures" with 91% accuracy*
- *n*-gram and Hidden Markov Models
  - linearize events into a sequence
- Hierarchical models of Life Stories
  - a Life Story is a (possibly overlapping) sequence of *careers*
  - each *career* is a sequence of *events*
  - each *event* has *properties* and a *duration*

Outline

FreeBase "notable types" as Life Story summaries

- Life Stories intuition: *background information about entities can improve language understanding*
  - goal: demonstrate that FreeBase's entity information improves an NLP application
- FreeBase entities are assigned *notable types*
  - examples: politician, celebrity, academic, cricket player, etc.
- Methodology:
  - take a state-of-the-art *relation extraction system*
    - Riedel, Yao, McCallum and Marlin (2013) *Relation Extraction with Matrix Factorization and Universal Schemas*
  - modify it to exploit notable types
  - measure the change in relation extraction accuracy

Relation extraction as matrix completion



- *Rows* encode *entity mid tuples* linked from New York Times text
  - named entities in text linked to FreeBase mids with Stanford NEL
- *Columns* encode *relations* in the text or in FreeBase
  - notable types *provide additional columns in the matrix*

The relation extraction model

- Relations are predicted using three different submodels (Riedel et al 2013)
  - *neighbourhood model:* predicts a matrix entry given the other entries along a row
  - *latent feature model:* associates each entity tuple (row) with a *100-dimensional latent feature vector*, and uses this to predict the relation
  - *entity-relation model:* a bilinear model associating entities with latent feature vectors and relations with latent matrices
- *Notable types* are added as additional latent feature vectors to each of these models

$\Rightarrow$ *A 10% overall error reduction over the state-of-the-art baseline*

Outline

Natural language understanding and text data mining

Named entity linking and relation extraction

"Life stories" models

Adding "notable types" to a relation extraction system

Conclusion and future work

MACQUARIE UNIVERSITY

Conclusion and future work

- The rise of mobile computing is driving the construction of large-scale knowledge bases such as FreeBase
- These large-scale knowledge bases can be exploited to improve natural language processing
  - "notable type" information about entities obtained from FreeBase reduces error in relation extraction by 10%
- The "Life Stories" project models the sequences of events that occur to individuals
  - being able to predict what events are likely to occur to an individual should dramatically improve a range of information extraction processes

Other applications of "Life Stories" models

- "Life Stories" models predict the future given a *structured past*
  - adds a *temporal dimension* to standard ML algorithms
- Other potential applications of "Life Stories" models:
  - *Finance:* identify corporate "success stories" and learn typical "failure modes"
  - *Health:* data mine medical records to learn disease trajectories and recovery pathways
  - *Insurance:* flag low-probability Life Stories as potential fraud
- "Life Stories" models are essentially models of *graph evolution*
  - FreeBase is itself just a gigantic labelled graph
  - A "Life Stories" model predicts how nodes (entities) and arcs (properties) are added to the graph