Life Stories

Mark Johnson
Wray Buntine, Lan Du, Anish Kumar
Massimiliano Ciaramita

Macquarie University
Sydney, Australia

March 2014

MACQUARIE UNIVERSITY

Which Jim Jones?

- News text: *Jim Jones' recent musical releases . . .*
- 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

What are Life Stories?

- FreeBase as a repository of Life Stories
  - FreeBase contains more than 100 properties for $\approx$ 250,000 people
  - Coverage is uneven: Sarah Palin's political career is covered, her political commentator roles on Fox News are not
- What appears in a Life Story?
  - time-stamped properties, e.g., *Bill Clinton's presidency 1993–2001*
  - indirectly time-stamped properties, e.g., *Bill Clinton's 1996 presidential campaign*
  - some properties without timestamps, e.g., *gender, nationality, notable type*
- Possible formalisations of Life Stories
  - temporally-bounded sets of events (i.e., a time-line)
  - events occuring in fixed windows (e.g., each year's events)

Important events

- Events differ in importance
  - Bill Clinton made 97 political appointments,
    appeared on 24 TV shows, and
    was elected US President twice
- FreeBase internal measures of importance (?)
  - *causes* are highly predictive, temporally-preceding event types (?)
- External measures of importance or impact
  - use relation extraction to align FreeBase properties to the individual's Wikipedia text, or a large news corpus
  - estimate importance by *amount of text* (sentences, column inches, etc.) linked to event

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 should be useful in
  - named entity linking
  - relation extraction

  but accuracy on those tasks depends on other factors as well
- Evaluate the predictive ability of a Life Story model, e.g.:
  - 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
- *n*-gram and Hidden Markov Models
  - linearize events into a sequence
  - project events onto a finite set of event types
- 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*

What's next

- We're currently preparing the data
- Next steps:
  - define evaluations
  - evaluate baseline models
  - develop better models
- We welcome suggestions and feedback!
- Can FreeBase improve a real information extraction task?
  - Anish Kumar's poster explains how *FreeBase's "notable types" improve a relation extraction system*

We're recruiting PhD students and post-docs!