Text data mining for information extraction

Mark Johnson

Macquarie University
Sydney, Australia

August 2014

MACQUARIE UNIVERSITY

Understanding natural language

- Understanding language is challenging because:
  - words *combine recursively* to form phrases and sentences with a *complex hierarchical structure*
  - which induce *non-local temporal dependencies* between the elements
- These techniques can *recover complex dependencies in other kinds of data* as well

Named Entity Recognition and Linking

$$\underbrace{\textit{Exoenzyme S}}_{\text{Protein}} \textit{ is an } \underbrace{\textit{extracellular product}}_{\text{Location}} \textit{ of } \underbrace{\textit{Pseudomonas aeruginosa}}_{\text{Organism}}$$

- *Named entity recognition* involves:
  - identifying words or phrases that refer to people, places or things
  - determining the *type of thing* referred to (e.g., company, disease)
- *Named entity linking* connects mentions to external databases:
  - e.g., drug names to drug databases, disease names to ICD codes
  - *ambiguity*: Wikipedia contains 6 “Anthony Abbott”s
- *Relation extraction* identifies “who did what to whom”:
  - converts unstructured text into a database format

Topic models and document clustering

- Topic models *simultaneously cluster* both documents and the words they contain:
  - documents are similiar if they contain similiar words
  - words are similiar if they appear in similiar documents
- Useful for *understanding very large data collections*
  - finds common themes or trends across the collection
  - identifies outliers that don't fit into any clusters
- Same techniques can be used to *analyse any database where records contain many recurring elements* (e.g., patient insurance records, financial transactions)
  - mathematically possible to *combine quantitative and qualitative information*

Example: documents from NIPS corpus

Annotating an unlabeled dataset is one of the bottlenecks in using supervised learning to build good predictive models. Getting a dataset labeled by experts can be expensive and time consuming. With the advent of crowdsourcing services . . .

The task of recovering intrinsic images is to separate a given input image into its material-dependent properties, known as reflectance or albedo, and its light-dependent properties, such as shading, shadows, specular highlights, . . .

In each trial of a standard visual short-term memory experiment, subjects are first presented with a display containing multiple items with simple features (e.g. colored squares) for a brief duration and then, after a delay interval, their memory for . . .

Many studies have uncovered evidence that visual cortex contains specialized regions involved in processing faces but not other object classes. Recent electrophysiology studies of cells in several of these specialized regions revealed that at least some . . .

Example (cont): ignore function words

Annotating an unlabeled dataset is one of the bottlenecks in using supervised learning to build good predictive models. Getting a dataset labeled by experts can be expensive and time consuming. With the advent of crowdsourcing services . . .

The task of recovering intrinsic images is to separate a given input image into its material-dependent properties, known as reflectance or albedo, and its light-dependent properties, such as shading, shadows, specular highlights, . . .

In each trial of a standard visual short-term memory experiment, subjects are first presented with a display containing multiple items with simple features (e.g. colored squares) for a brief duration and then, after a delay interval, their memory for . . .

Many studies have uncovered evidence that visual cortex contains specialized regions involved in processing faces but not other object classes. Recent electrophysiology studies of cells in several of these specialized regions revealed that at least some . . .

Example (cont): mixture topic model

Annotating an unlabeled dataset is one of the bottlenecks in using supervised learning to build good predictive models. Getting a dataset labeled by experts can be expensive and time consuming. With the advent of crowdsourcing services . . .

The task of recovering intrinsic images is to separate a given input image into its material-dependent properties, known as reflectance or albedo, and its light-dependent properties, such as shading, shadows, specular highlights, . . .

In each trial of a standard visual short-term memory experiment, subjects are first presented with a display containing multiple items with simple features (e.g. colored squares) for a brief duration and then, after a delay interval, their memory for . . .

Many studies have uncovered evidence that visual cortex contains specialized regions involved in processing faces but not other object classes. Recent electrophysiology studies of cells in several of these specialized regions revealed that at least some . . .

Example (cont): admixture topic model

Annotating an unlabeled dataset is one of the bottlenecks in using supervised learning to build good predictive models. Getting a dataset labeled by experts can be expensive and time consuming. With the advent of crowdsourcing services . . .

The task of recovering intrinsic images is to separate a given input image into its material-dependent properties, known as reflectance or albedo, and its light-dependent properties, such as shading, shadows, specular highlights, . . .

In each trial of a standard visual short-term memory experiment, subjects are first presented with a display containing multiple items with simple features (e.g. colored squares) for a brief duration and then, after a delay interval, their memory for . . .

Many studies have uncovered evidence that visual cortex contains specialized regions involved in processing faces but not other object classes. Recent electrophysiology studies of cells in several of these specialized regions revealed that at least some . . .

Our innovation: topical multi-word expressions

Annotating an unlabeled dataset is one of the bottlenecks in using supervised learning to build good predictive models. Getting a dataset labeled by experts can be expensive and time consuming. With the advent of crowdsourcing services . . .

The task of recovering intrinsic images is to separate a given input image into its material-dependent properties, known as reflectance or albedo, and its light-dependent properties, such as shading, shadows, specular highlights, . . .

In each trial of a standard visual short-term memory experiment, subjects are first presented with a display containing multiple items with simple features (e.g. colored squares) for a brief duration and then, after a delay interval, their memory for . . .

Many studies have uncovered evidence that visual cortex contains specialized regions involved in processing faces but not other object classes. Recent electrophysiology studies of cells in several of these specialized regions revealed that at least some . . .

MACQUARIE
UNIVERSITY

“Life stories”: uncovering temporal structure

- “Life story” models capture the temporal structure of events
    - a *life story* is the sequence of events that occur to a person
    - a *career* is a highly-correlated sequence of events
    - any individual’s life story involves multiple careers
- We learn typical careers from large numbers of life stories
- and use these models to *predict likely future events* from a partial life story

Summary

- Natural language understanding involves identifying complex temporal and structural patterns
- We can automatically identify named entities in text and link them to databases
- Topic models jointly cluster “documents” and the “words” they contain
  - identify common trends and outliers
- Life story models generalise topic models by learning a temporal structure to topics