Natural Language Processing from a Machine Learning Perspective

Mark Johnson

November 2015

What is Natural Language Processing?

- *Natural Language Processing* (NLP) develops methods for *transforming* or *extracting information* from *text or speech*
- Classic examples of NLP:
  - machine translation
  - speech recognition (automatic transcription)
  - summarisation (single document or multi-document)
  - *human-computer interaction* (e.g., question-answering)

A short history of Natural Language Processing

- Machine translation started with the first computers in the 1950s
  - deeply influenced by the Chomskyian *cognitive revolution*
- Until the 1990s NLP centred around:
  - implementing linguistic theories of syntax
  - with parsers based on symbolic AI theorem-proving methods
- The *statistical revolution* started in speech recognition
  - Hidden Markov Models worked better than rule-based systems
  - in general, probabilistic approaches work better than rule-based ones
- We may be at the start of a *deep learning neural network revolution*

Outline

Brief review of machine learning

Document classification and sentiment analysis

Named entity recognition and linking

Syntactic parsing and relation extraction

Topic modeling

Conclusions and future directions

Prediction vs. causation

- Classical statistics focuses on discovering *causal relationships*
  - E.g., *does coffee cause lung cancer?*
  - it's hard to identify causal dependencies between more than ≈ 10 variables
- Machine learning and data mining focus on *prediction*
  - E.g., *how many people are likely to get lung cancer?*
  - variables can have predictive value even if the causal dependencies aren't clear
  - E.g., *maybe smoke in coffee-houses is to blame?*

$\Rightarrow$ *can learn predictive models with millions of variables*

Supervised vs. unsupervised learning

- *Prediction problems* use data D to predict the value of a variable y from other variables x
  - E.g., x = *a patient's medical test results today*
  - y = *whether they have lung cancer 5 years from now*
  - D = *other patients' medical results from 5 years ago, and their current lung cancer status*
- In *supervised learning* the data D contains the variable y we want to predict
- In *unsupervised learning* the data D does not contain the variable y we want to predict
- There is a continuum between supervised and unsupervised learning, including:
  - *semi-supervised learning:* only some of the data is labeled
  - *distant supervision:* D is labeled with a variable related to y
  - *domain adaptation:* D comes from a different population

A typology of machine learning problems

- The nature of the predicted or dependent variable y determines the kind of problem and algorithm involved
  - y can be *categorical* or *discrete*, e.g., *is the patient alive?*
  - y can be *continuous*,
    e.g., *what is the patient's lung capacity?*
- Mapping problems to algorithms:

	discrete y	*continuous y*
supervised data	*classification*	*regression*
unsupervised data	*clustering*	*dimensionality reduction*

Outline

Document classification and bag-of-words features

- In *document classification*, x is a document (e.g., a news story) and y is e.g., *sports/finance/current affairs*
- A good baseline model treats x as an unordered *bag of words*, i.e., a vector with a dimension for each word in the vocabulary

  A man who allegedly tried to run over a police officer before speeding off has been arrested at a Melbourne police station after turning up in a stolen car carrying guns and drugs.

$$[\underbrace{3}_{\text{a}}, \underbrace{1}_{\text{man}}, \underbrace{0}_{\text{woman}}, \underbrace{1}_{\text{allegedly}}, \underbrace{0}_{\text{alleged}}, \underbrace{0}_{\text{try}}, \underbrace{1}_{\text{tried}}, \ldots]$$

- Standard regression and classification algorithms (e.g., SVMs) work well with bag-of-words representations, so long as they use *sparse vector* techniques to handle the large number of features (vocabulary size $> 10,000$)

Sentiment analysis and opinion mining

- Sentiment analysis and opinion mining is a commercially-important application of document classification
  - typical application: social media posts
- Usually consists of two classifiers:
  - Classifier 1 classifies documents as objective/sentimental
  - Classifier 2 classifies documents as $+/-$ sentiment
- Bag-of-words representation works well for sentiment analysis of restaurant reviews, but badly for movie reviews
  - E.g., *I liked the start of the movie, but towards the middle I started to get bored . . .*
  - modeling syntactic and discourse structure greatly improves sentiment analysis of movie reviews
- *Aspect-based sentiment analysis* associates sentiment with entities mentioned in the document

Outline

Brief review of machine learning

Document classification and sentiment analysis

Named entity recognition and linking

Syntactic parsing and relation extraction

Topic modeling

Conclusions and future directions

Named entity recognition and linking

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

$$\underbrace{\textit{Malcolm Turnbull}}_{\text{person}}\ \textit{bought}\ \underbrace{\textit{300}}_{\text{number}}\ \textit{shares in}\ \underbrace{\textit{Acme Corp}}_{\text{corporation}}\ \textit{in}\ \underbrace{\textit{2006}}_{\text{date}}$$

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

  Example: *Malcolm Turnbull met the president of Indonesia yesterday. Mr. Turnbull told him that he . . .*

- *Entity linking* maps entities to database entries

$$\underbrace{\textit{Malcolm Turnbull}}_{\text{/m/xw2135}}\ \textit{bought}\ \underbrace{\textit{300}}_{\text{number}}\ \textit{shares in}\ \underbrace{\textit{Acme Corp}}_{\text{/m/yzw9w}}\ \textit{in}\ \underbrace{\textit{2006}}_{\text{date}}$$

Sequence labelling problems

- A *sequence labelling* problem is one where:
  - the input consists of a sequence $\boldsymbol{X} = (X_1, \ldots, X_n)$, and
  - the output consists of a sequence $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ of labels, where:
  - Y_i is the label for element X_i
- Example: Part-of-speech tagging

$$\begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{X} \end{pmatrix} = \begin{pmatrix} \text{Verb,} & \text{Determiner,} & \text{Noun} \\ \text{spread,} & \text{the,} & \text{butter} \end{pmatrix}$$

- Example: Spelling correction

$$\begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{X} \end{pmatrix} = \begin{pmatrix} \text{write,} & \text{a,} & \text{book} \\ \text{rite,} & \text{a,} & \text{buk} \end{pmatrix}$$

Named entity extraction as sequence labelling

- NER can be formulated as a sequence labelling problem by using the *Inside-Outside-Begin* (IOB) labelling scheme

B-ORG	I-ORG	O	O	O	B-LOC	I-LOC	I-LOC	O
Macquarie	University	is	located	in	New	South	Wales	.

- The IOB labelling scheme can distinguish *adjacent named entities*

B-LOC	I-LOC	I-LOC	B-LOC	I-LOC	O	B-LOC	O
New	South	Wales	Northern	Territory	and	Queensland	are

Other applications of sequence labelling

- *Speech recognition* is a sequence labelling task:
  - The input $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a sequence of *acoustic frames* X_i, where X_i is a set of features extracted from a 50msec window of the speech signal
  - The output $\boldsymbol{Y}$ is a sequence of words (the transcript of the speech signal)
- Financial applications of sequence labelling
  - identifying trends in price movements
- Biological applications of sequence labelling
  - gene-finding in DNA or RNA sequences

A first (bad) approach to sequence labelling

- Idea: train a supervised classifier to *predict entire label sequence at once*

B-ORG	I-ORG	O	O	O	B-LOC	I-LOC	I-LOC	O
Macquarie	University	is	located	in	New	South	Wales	.

- Problem: *the number of possible label sequences grows exponentially with the length of the sequence*
  - with *binary labels*, there are 2^n different label sequences of a sequence of length n ($2^{32} = 4$ billion)

$\Rightarrow$ most labels won't be observed even in very large training data sets

- This approach fails because it has massive *sparse data problems*

A better approach to sequence labelling

- Idea: train a supervised classifier to *predict the label of one word at a time* (slide a "moving window" over the text)

B-LOC	I-LOC	O	O	O	O	O	B-LOC	O
Western	Australia	is	the	largest	state	in	Australia	.

- Avoids sparse data problems in label space
- As well as current word, classifiers can use *previous and following words as features*
- But this approach can produce *inconsistent label sequences*

O	B-LOC	I-ORG	I-ORG	O	O	O	O
The	New	York	Times	is	a	newspaper	.

⇒ Track *dependencies between adjacent labels*

  - "chicken-and-egg" problem that *Hidden Markov Models* and *Conditional Random Fields* solve!

Outline

Brief review of machine learning

Document classification and sentiment analysis

Named entity recognition and linking

Syntactic parsing and relation extraction

Topic modeling

Conclusions and future directions

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

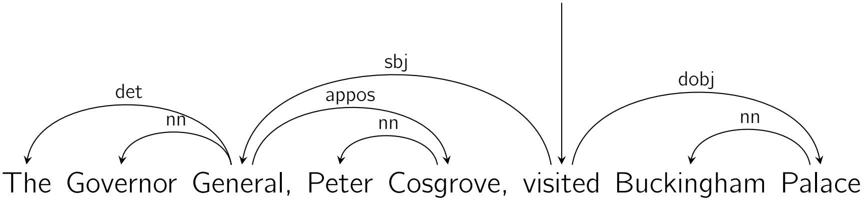
Person	**Role**
Peter Cosgrove	Governor General of Australia

  Offical-visit

Visitor	**Organisation**
Peter Cosgrove	Queen of England

- Text-mining *bio-medical literature* is a major application

Syntactic parsing for relation extraction

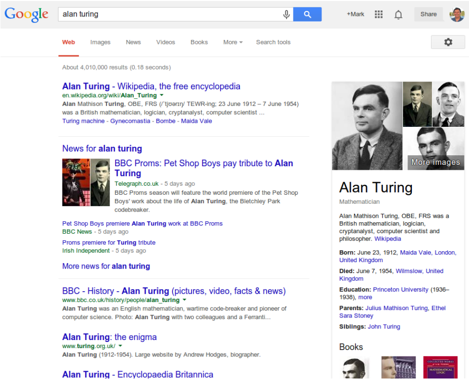


- The *syntactic path* in a *dependency parse* is a useful feature in relation extraction

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

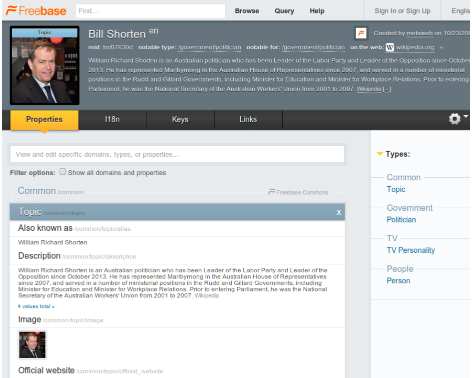
$$X \xleftarrow{\text{sbj}} \textit{visited} \xrightarrow{\text{dobj}} Y \Rightarrow \textit{official-visit}(X, Y)$$

Google's Knowledge Graph



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

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, 32GB compressed dump
- Created by Metaweb, which was acquired by Google

Distant supervision for relation extraction

- Ideal labelled data for relation extraction: large text corpus annotated with entities and relations
  - expensive to produce, especially for a lot of relations!
- *Distant supervision assumption:* if two or more entities that appear in the same sentence also appear in the same database relation, then probably the sentence expresses the relation
  - assumes entity tuples are sparse
- With the distant supervision assumption, we obtain relation extraction training data by:
  - taking a large text corpus (e.g., 10 years of news articles)
  - running a named entity linker on the corpus
  - looking up the entity tuples that appear in the same sentence in the large knowledge base (e.g., FreeBase)

Outline

Brief review of machine learning

Document classification and sentiment analysis

Named entity recognition and linking

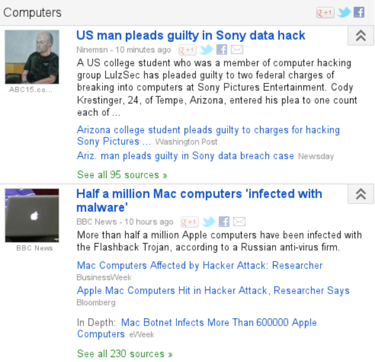
Syntactic parsing and relation extraction

Topic modeling

Conclusions and future directions

Topic modelling

- Topic models *cluster words and documents into topics*
  - *unsupervised* (i.e., topics aren't given in training data)
- Important for document analysis and information extraction
  - Example: clustering news stories for information retrieval
  - Example: tracking evolution of a research topic over time



Mixture versus admixture topic models

- In a *mixture model*, each document has a *single topic*
  - all words in the document come from this topic
- In *admixture models*, each document has a *distribution over topics*
  - a single document can have multiple topics (number of topics in a document controlled by prior)
  - $\Rightarrow$ can capture more complex relationships between documents than a mixture model
- Both mixture and admixture topic models typically use a *"bag of words"* representation of a document

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: Collocation topic models

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 other work on topic models

- *Segment documents into topically-coherent parts:* find major topic shifts in an unsegmented document (e.g., speech recogniser output)
- *Integrate topic modelling with other information:* improve topic model accuracy by using additional information, e.g., social follower information, sentiment, etc.

Outline

Brief review of machine learning

Document classification and sentiment analysis

Named entity recognition and linking

Syntactic parsing and relation extraction

Topic modeling

Conclusions and future directions

Overview and summary

- Current NLP technology does not understand language the way people do, but it can work fairly well
- Simple "bag of words" methods are often surprisingly effective on some document genres
  - *document classification* accuracy varies depending on genre and information you want to extract (70% to 90% is typical)
  - *topic models* are an unsupervised approach that clusters words and documents
- Sequence models and syntactic parsing models identify relationships between words
  - important for identifying *who did what to whom?*

Directions for future work

- The probabilistic models and statistical methods underlying NLP are the same as those used in data analytics

⇒ Combine *data analytics of structured data* with *text data mining* of unstructured data
  - E.g., structured data: medical test results, purchase history, etc.
    unstructured data: medical records, social media posts, etc.

- The techniques that find *named entities* in texts should be able to mine numerical quantities, dates, currency amounts, etc., in unstructured text
  - integrating these in a *joint model* should improve text data mining and data analytics