Building an international database of child policy and governance in education

Robert Goerge
Chapin Hall at the University of Chicago

Dominic Richardson
UNICEF Office of Research - Innocenti

ISCI Conference
September 2, 2015
Cape Town, South Africa

Overview

- A major challenge for international monitoring of child well-being is compiling comprehensive information on the policies and practices of countries and sub-country units in domains of child wellbeing.
- Typically, this is a very labor-intensive activity, which requires officials from countries completing surveys or researchers searching for documents or websites which contain the policies.
- Currently, collecting such information is time-intensive and, therefore, probably not done as often as it should be
- The intention of this project is for use computer programs to search on-line (internet), **public** resources of information about education policies in countries across the globe – ***essentially to look through a large amount of on-line documents and extract meaning from them.***

Text Mining and Natural Language Processing Examples

- Search engines
- Google Trends, flu trends
  - http://www.google.com/trends/
  - http://www.google.org/flutrends/
- Product review mining
  - e.g., Amazon.com
- Education example
  - http://www.gov.uk/government/collections/statutory-guidance-schools
  - http://www.theguardian.com/politics/education

Text mining of text in government databases

- More and more text is being added to government databases.
- It used to be collected on paper in filing cabinets.
- This text data can be mined to provide data from unstructured, non-numeric data and analyzed.

How things are done now

Desk-based search of online resources:

- Existing international databases (MISSOC, WORLD etc.)
- Supplemented via government websites

Questionnaires sent to government ministries

- Sometimes pre-filled

Repeated annually

Text-mining

- Text-mining involves accessing unstructured text from online or other sources and using computer programming to access data about specific policy data. This requires identifying word, patterns of works, a dictionary of terms and synonyms.
- Obviously, it requires ultimately doing this work in languages other than English, although much of this information across the world is available online in English.
- Once the program is written for a particular country, it can be modified and re-run in minutes to obtain updated data.

Steps in the text mining process

- Text mining is the process of mechanically finding non-trivial patterns from unstructured text data.
- A software program finds and flags terms that appear several times within some text data, that is also a type of text mining.
- A computer program “comprehends” text by mapping words and phrases to concepts.
- By mapping words and phrases to concepts, text mining creates structured data from unstructured; each ***x*** concept is connected to a ***y*** word or phrase by the text mining program. Once patterns have been noted and the data has been organized into a structured format, researchers can analyze the data in various ways.
- As textual sources continue to become more plentiful, increasingly sophisticated and powerful text mining methods and tools are being developed and used for a variety of purposes. The tools are often able to efficiently process large amounts of data, in real--time or as batch jobs, and can scale to accommodate increased demands.

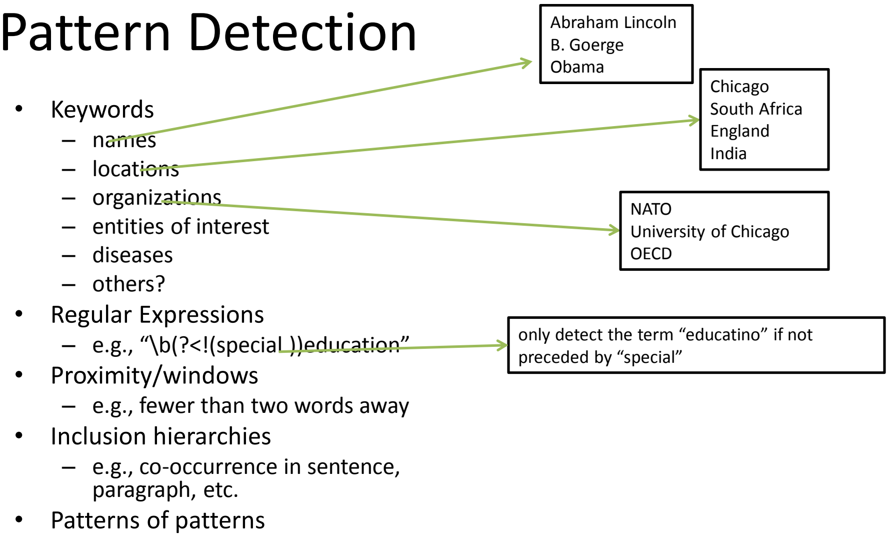
Text mining of government policy

- To facilitate transparency and informing the public of polices, regulations and procedures, governments are creating websites that contain this information
- These data are not typically “coded” into structured or numeric variables that can be analyzed with quantitative methods. They are text documents with potentially a different format in each country.
- One has to find where the relevant documents exist to be able to extract the information that is needed.

Natural Language Processing

- Pattern detection
  - Look for known patterns
- Machine Learning
  - Learn based on text coded by experts
  - Learn without external input
- Often a hybrid approach is used
  - Combine the best of both worlds

Pattern Detection



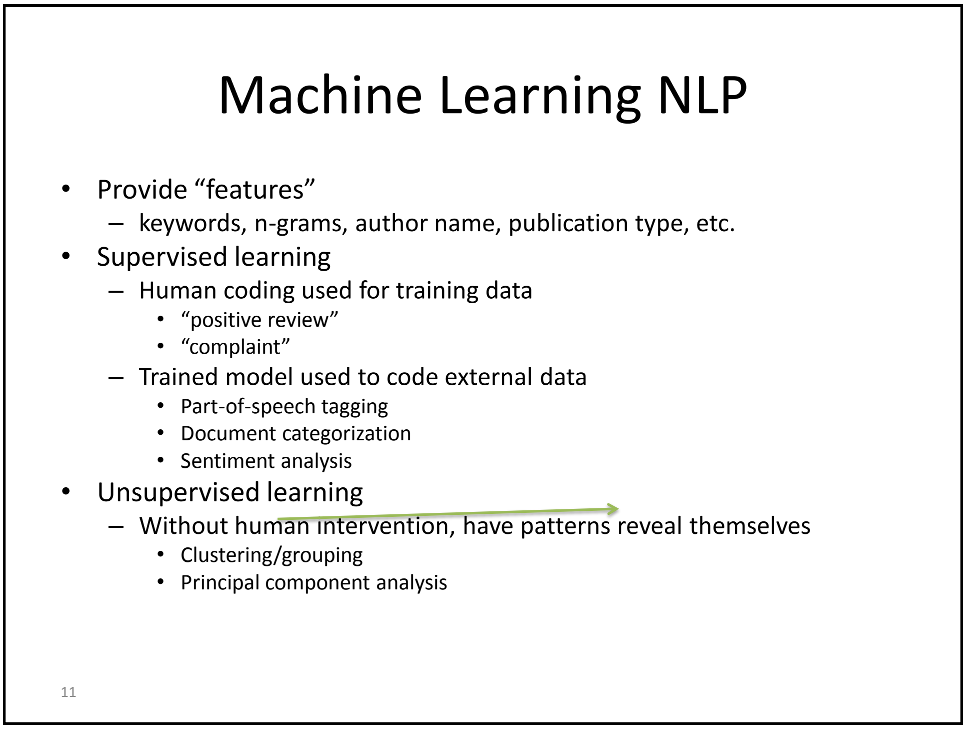
- Keywords
  - names
  - locations
  - organizations
  - entities of interest
  - diseases
  - others?
- Regular Expressions
  - e.g., “\b(?<!(special))education”
- Proximity/windows
  - e.g., fewer than two words away
- Inclusion hierarchies
  - e.g., co-occurrence in sentence, paragraph, etc.
- Patterns of patterns

Abraham Lincoln
B. Goerge
Obama

Chicago
South Africa
England
India

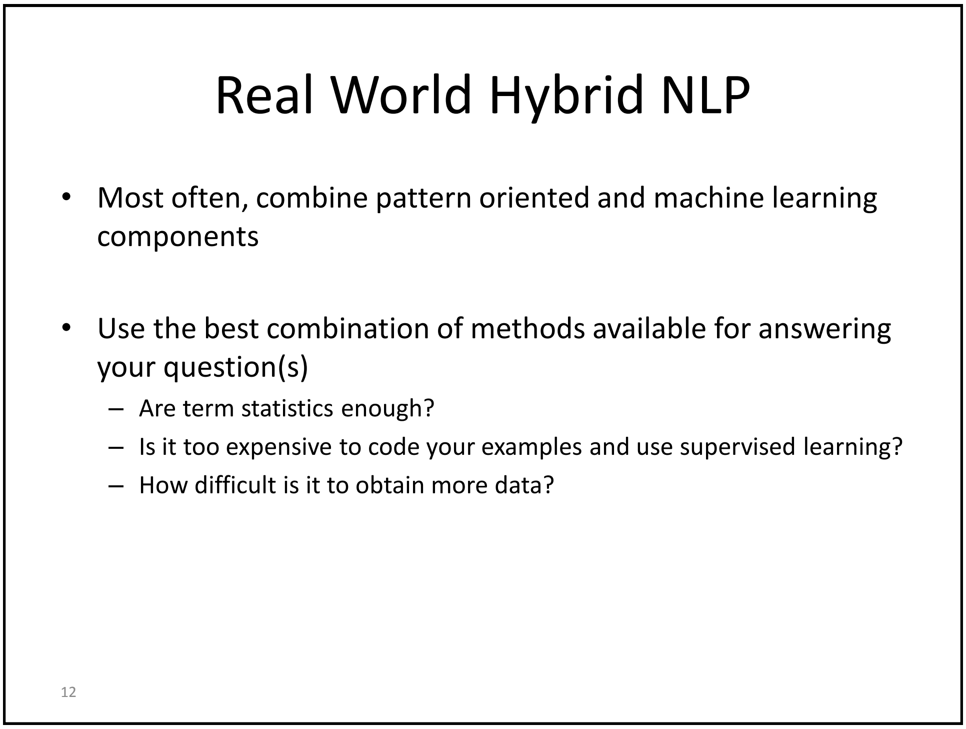
NATO
University of Chicago
OECD

only detect the term “educatino” if not preceded by “special”



Machine Learning NLP

- Provide “features”
  - keywords, n-grams, author name, publication type, etc.
- Supervised learning
  - Human coding used for training data
    - “positive review”
    - “complaint”
  - Trained model used to code external data
    - Part-of-speech tagging
    - Document categorization
    - Sentiment analysis
- Unsupervised learning
  - Without human intervention, have patterns reveal themselves
    - Clustering/grouping
    - Principal component analysis



Real World Hybrid NLP

- Most often, combine pattern oriented and machine learning components

- Use the best combination of methods available for answering your question(s)
  - Are term statistics enough?
  - Is it too expensive to code your examples and use supervised learning?
  - How difficult is it to obtain more data?

Intricacies - Sentiment Analysis

- What should we expect to detect?

via Twitter → iPhone > Android > Nokia > Land phone > Typewriter > 2 cans and a string > Message in a bottle > Pigeon with a note > Blackberry

via Amazon →

I LOVE APPLE but lukewarm on iPhone 4S

I love my Apple Products and will continue to buy them. However, am a little disappointed in the new 4S. The mapping features have a lot of inaccuracies. Siri can sometimes help, but is mostly annoying. She stopped being funny a long time ago. I have four kids that can get smart with me. If she was more helpful, it wouldn't be as annoying. But, in all honesty, I find it easier and less distracting to simply use the internet. An important footnote. If you are new to the iPhone, I think getting the 4S is a good idea. It is the best of the iPhones. But if you have an iPhone 4 (like I did), you might just want to wait until the next generation!

Public Laws

- http://www.archives.gov/federal-register/laws/current.html

Regulations

- http://www.dhs.state.il.us/page.aspx?Item=13473

Intricacies - Getting Data

- Real time or near real time
  - Twitter API (https://dev.twitter.com/), TwitterSearch (https://pypi.python.org/pypi/TwitterSearch/)
- News
  - e.g., NYTimes API (http://developer.nytimes.com/)
- Web scraping
  - e.g., Apache Tika (http://tika.apache.org/), lxml (http://lxml.
- Web crawling
  - e.g., Apache Nutch (http://nutch.apache.org/), Scrapy (http://scrapy.org)
- Text archives
- The difficulties in getting enough high qual hould not be overlooked…