Some things I'll talk about

- What is Natural Language Processing?
- Where does it fit into search?
  - Question answering
  - Machine translation
  - Summarization
- Academic approaches and real-life examples from Google
What is NLP?

- Branch of AI focused on making computers more useful for people "on our own terms"
- Includes research in areas like
  - Search*
  - Machine translation*
  - Speech recognition
- Useful interaction != Sentient understanding
Search must expand with the web

- Beyond simple, static text documents
- "Universal search" blends result types - how to rank?
How can NLP improve search?

- Understanding queries
  - [who is the fordham law school dean?]
- Understanding documents
  - Mining clean "facts" from messy web documents
- Finding and presenting useful results
  - Via translation and/or summarization
- Not getting in the way
  - Don't need deep NLP for [home depot]
- Rest of this talk: where is NLP both useful and practical?
How much NLP do we want, then?

- **Tokenization**
  - \textit{seek} \neq \textit{seeking}
  - Easy in English, not as easy in Chinese

- **Stemming / morphology**
  - \textit{seek} \approx \textit{seeking} \approx \textit{sought}
  - Rule-based approaches (e.g., Porter) cover many, not all cases

- **Synonyms and translation**
  - \textit{seek} \approx \textit{search}
  - Manual (e.g., WordNet) and automatic (e.g., LSA) techniques
  - Via machine translation (MT), \textit{seek} \approx \textit{cherchez}

- **Named entities and frame semantics**
  - Frames as context, e.g. \textbf{TEACHING} (teacher, student, subject)
  - Named entities and types

- **Topic and discourse structure**
  - Discourse relations, e.g. \textbf{CAUSE}, \textbf{CONTRAST}
  - Summarizing most "important" information
Named entities

- Named entities (NEs) =~ natural language "types"
  - e.g., PERSON, LOCATION, DATE
  - cf., Sundheim (1991) and MUC conferences
- For question answering
  - Semantic frames can combine with NEs to verify if an answer "makes sense"
  - e.g., a PERSON was born on a DATE
- For summarization
  - We can form a priori rules around NEs
  - Expectation that PERSON or PLACE information has high informational value
Detecting named entities

- Rule/dictionary-based methods
  - Coverage issues (especially multilingual)
  - Ambiguity (e.g. is Paris PERSON or LOCATION?)
- Statistical methods, e.g. supervised training
  - If we see, ... traveled to Paris for vacation ...
  - We can determine, e.g., maximum likelihood:
    - $P(\text{LOCATION}(t_n) | t_n == \text{Paris}) = X$
    - $P(\text{LOCATION}(t_n) | t_{n-2},t_{n-1} == \text{traveled to}) = Y$
    - Naive bayes estimate = $X \times Y$
- Web-scale techniques
  - Must be robust, internationalizable
  - More training material available
  - Bootstrapping methods important
  - cf., Callan and Mitamura (2002)
Using named entities as a "view"

- Entity detection can enable search on different dimensions.
Entities and question answering (QA)

- Some queries are fact-seeking questions
  - [who is the fordham law school dean?]
  - [new zealand population]
  - [find me a cheap, public golf course with no water hazards near st. johnsbury vermont]
- For Google QA, database of facts about entities
  - Mined from web documents
  - Searched when you enter a query, e.g.
    - New Zealand: population = 4.17M
    - Shaquille O'Neal: height = 7'1"
    - Fordham Law: dean = William Treanor
Mining facts: information extraction

- Information extraction (IE)
  - General technique of finding structured data in text or other unstructured sources (cf., Grishman (1997))
  - Continuum of manual -> automatic techniques
- Scrapers for individual sites
  - URL: https://www.cia.gov/cia/publications/factbook/*
  - Population: “Population:.*<br>([{0-9,}]*)”
- Patterns for text extraction
  - “the population of ([a-z]+) is ([0-9,]+)”
- Scaling issues!
"Bootstrapping" leverages unannotated data
- Input small number of manual "seed" facts
  - Vincenzo Bellini: birthdate: 1801
  - Bob Dylan: birthdate: 1941
- Find instances of seeds in text corpus (NEs help)
  - Vincenzo Bellini was born in Catonia in 1801
  - Bob Dylan (1941 - present) is best known ...
- Generalize patterns from seed instances
- Repeat to refine patterns, add new seed facts, ...

- Not just person-birthdate, but company-headquarters, country-president, etc.
- For Google, ~1M facts from ~100M docs
IE Bootstrapping in Snowball

● Well-known IE system (Agichtein and Gravano, 2000)
● If we have two facts which are (PERSON, YEAR) pairs:
  ○ Vincenzo Bellini: birthdate: 1801
  ○ Bob Dylan: birthdate: 1941
● Find instances in the corpus, and record contexts:
  ○ [and] Vincenzo Bellini [ was born in Catonia in ] 1801 [ . ]
  ○ [ . ] Bob Dylan [ was born Robert Zimmerman in ] 1941 [ to ]
● Learns flexible patterns as Left/Middle/Right vectors
  ○ L = {(., 0.5), (and , 0.5)}
  ○ M = {((born , 1.0), (was , 1.0), (in , 1.0), (Catonia , 0.5) ...}
  ○ R = {(., 0.5), (to , 0.5)}
● Use patterns to find and rank new seeds, repeat
Machine translation and CLIR

- Machine translation (MT) translates natural language documents.
- In the search context, we first need to find relevant documents.
  - As a speaker of language X, I don't necessarily know how to write queries in language Y.
- Cross-language information retrieval (CLIR) addresses this issue via translate-search-translate.
  - [Identifying wild mushrooms] in French!
MT: How does it work?

- Collect parallel texts
  - i.e., texts that have been translated by human translator
  - texts from multilingual governing bodies, e.g. UN, are often used

DE

SEHR GEEHRTER GAST! KUNST, KULTUR UND KOMFORT IM HERZEN BERLIN.

EN

DEAR GUESTS, ART, CULTURE AND LUXURY IN THE HEART OF BERLIN.
MT: How does it work (II)

KUNST, KULTUR UND KOMFORT IM HERZEN BERLINS.

|   |   |   |   |   |   |   |   / | / | \ |   | / | \ |   | \ |

ART, CULTURE AND LUXURY IN THE HEART OF BERLIN.

- Align texts, learning probabilities for
  - Fertility: *berlins* -> *of berlin*
  - Translation: *kultur* -> *culture*
  - Distortion: *e.g.*, *green witch* -> *bruja verde*
- IBM model (Brown, Cocke et al., 1990)
  - Issues and improvements since then: many-to-one, distortion costs, syntax-aware
MT: How does it work (III)

\[ \hat{e}(f) = \arg\max_e Pr(e|f) \]
\[ = \arg\max_e \frac{Pr(e) \cdot Pr(f|e)}{Pr(f)} \]
\[ = \arg\max_e Pr(e) \cdot Pr(f|e) \]
\[ \approx \arg\max_e p(e) \cdot p(f|e) \]

- \( e \) = English, \( f \) = French
- "Solve for most likely English generating this French"
- Two key parts
  - Faithfulness: translation model: \( Pr(f|e) \)
  - Fluency: language model: \( Pr(e) \)
MT: More data helps ...  

Arabic to English, Five-gram language model, no count-cutoff, integrated into search
MT: … but more data is expensive!

- How many bits to store probabilities?
  - Double: 64 bit?
  - Float: 32 bit? (for trillions of numbers)
  - 16, 8, …? (8GB vs. 1GB)
- Use minimum size that retains acceptable performance!
MT: Price-performance tradeoffs

(for language model)
Text Summarization

- Capturing "gist" of document
- Various flavors
  - Single document vs multi-document
  - Query-focused vs document(s) only
  - Extractive: select verbatim text from document(s)
  - Abstractive: merge or synthesize new text
- Google search results are a summary!
  - query-focused, multi-document, extractive
Summarization techniques

- **Leading text**
  - Good baseline; formatting issues with web docs
- **Centroid** (Radev et al., 2000)
  - Documents in vector space, extract nearest to centroid
- **Graph-based** (Erkan and Radev, 2004)
  - Rank sentences by graph properties like connectedness
- **Entity/Event-focused** (Filatova and Hatzivassiloglou, 2004)
  - "Cover" major NE pairs and relationships
- **Speech features** (Maskey and Hirschberg, 2005)
  - For audio broadcasts, use features like emphasis
- **Sentence compression** (cf. McDonald 2006)
  - Under space constraints, drop relative clauses etc.
- **Opinion-focused** (cf. Pang and Lee 2008)
  - Summarizing over reviews, blogs etc.
Summarization in practice

- **Basic search results**: uses "snippets" as mini-summaries
- **Google news**: Story "clusters" summarize evolving events
- **Local search**: Snippets must be representative
Opinion summarization overview

- Important for local, product searches
- What is the polarity of the comments -- positive, negative?
  - Some reviews have "stars" -- but not for each sentence
- What aspects are people talking about?
  - General aspects, e.g. "service" or "value"
  - Specific aspects, e.g. "swimming pool" for a hotel, or "battery life" for a camera
Aspect-level opinion summarization

- **Sentiment Classification**
  - Identify all sentiment laden text in reviews.

- **Aspect Extraction**
  - Identify relevant aspects in these fragments.

- **Aspect Summarization**
  - Summarize aspects by sentiment

---

**What they like (or don't):**

- pizza (48.1%)
- food (55.7%)
- service (45.8%)
- atmosphere (90.0%)
- prices (67.3%)
- crust (58.3%)
- general (41.2%)

**service** - 177 total comments (81 positive, 42 neutral, 54 negative)

The service was poor all night long. ... I was willing to tolerate mediocre food if the service had been acceptable. ... Their service was great. ... The only enjoyable thing was the wine, and the service was ok. ...
Opinion lexicon learning

- Bootstrapping using opinion "seed" words
- Follow synonym/similarity links over word graph
  - Use label propagation to multiply / attenuate weights along edges (Zhu et al., 2002)
- Graph can be manually (e.g. WordNet) or automatically calculated (cf. Rao and Ravichandran, 2008)
  - Resource like WordNet provide more structure, e.g. antonym links, but do not scale/internationalize
Wrap-up

- Search is expanding
- Many challenges, opportunities for NLP
  - Question answering
  - Translation
  - Summarization
  - And more: opinion analysis, search with audio queries/documents, text classification, …
Questions?