CS6200
Information Retrieval

Jesse Anderton
College of Computer and Information Science
Northeastern University

Based largely on chapters from Speech and Language Processing, by Jurafsky and Martin

Information Extraction

- So far, we have focused mainly on ad-hoc web search. This usually starts from a user query and tries to find relevant documents.

- Another possible approach to IR is to construct a database of facts inferred from online text. This database can be used to answer questions more directly.

- The related task of Question Answering involves responding to a query with a textual answer instead of a list of documents.

Named Entity Recognition

Named Entity Recognition | Relation Extraction
Question Answering | Summarization

Named Entity Recognition

- **Named Entity Recognition** is identifying clauses in text which correspond to particular people, places, organizations, etc.

- Clauses are annotated with labels from a predefined list, such as:

Tag	Entity	Example
PER	People	Pres. Obama
ORG	Organization	Microsoft
LOC	Location	Adriatic Sea
GPE	Geo-political	Mumbai
FAC	Facility	Shea Stadium
VEH	Vehicles	Honda

Citing high fuel prices, **[ORG United Airlines]** said **[TIME Friday]** it has increased fares by **[MONEY $6]** per round trip on flights to some cities also served by lower-cost carriers. **[ORG American Airlines]**, a unit of **[ORG AMR Corp.]**, immediately matched the move, spokesman **[PERS Tim Wagner]** said.

Example of NER

Ambiguity in NER

- NER systems are faced with two types of ambiguity:
    - **Reference resolution**: the same name can refer to different entities of the same type. For instance, JFK can refer to a former US president or his son.
    - **Cross-type Confusion**: the identical entity mentions can refer to entities of different types. For instance, JFK also names an airport, several schools, bridges, etc.

JFK?

PRESIDENT KENNEDY
DEATH BY ASSASSIN
SHOT IN DALLAS
KEVIN COSTNER
JFK

NER as Sequence Labeling

- **Sequence labeling** is a common approach to NER.
- Tokens are labeled as:
    - B: Beginning of an entity
    - I: Inside an entity
    - O: Outside an entity
- We train a Machine Learning model on a variety of text features to accomplish this.

Word	Label
American	B
Airlines	I
a	O
unit	O
of	O
AMR	B
Corp.	I
immediately	O
matched	O
the	O
move	O
spokesman	O
Tim	B
Wagner	I
said	O

NER Features

Feature Type	Explanation
Lexical Items	The token to be labeled
Stemmed Lexical Items	Stemmed version of the token
Shape	The orthographic pattern of the word (e.g. case)
Character Affixes	Character-level affixes of the target and surrounding words
Part of Speech	Part of speech of the word
Syntactic Chunk Labels	Base-phrase chunk label
Gazetteer or name list	Presence of the word in one or more named entity lists
Predictive Token(s)	Presence of predictive words in surrounding text
Bag of words/ngrams	Words and/or ngrams in the surrounding text

NER Features

- In English, the shape feature is one of the most predictive of entity names.
- It is particularly useful for identifying businesses and products like Yahoo!, eBay, or iMac.
- Shape is also a strong predictor of certain technical terms, such as gene names.

Shape	Example
Lower	cummings
Capitalized	Washington
All caps	IRA
Mixed case	eBay
Capitalized character w. period	H.
Ends in digit	A9
Contains hyphen	H-P

Sequence Labeling

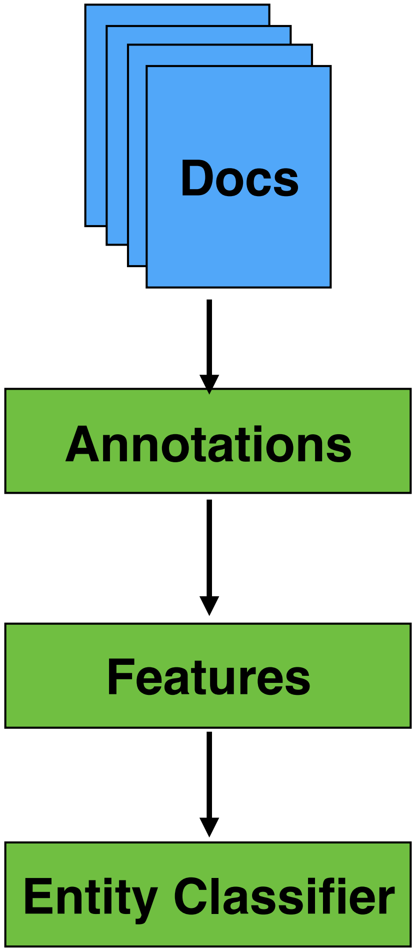
Steps of the sequence labeling process:

1. A collection of training documents is built

2. Humans annotate some or all of the entities in the training documents

3. Features are extracted

4. Classifiers are trained for each entity type



Training an IOB Encoder

- We train the IOB encoder using a slight variation of the standard classification task in Machine Learning.

- We plan to assign IOB tags sequentially to each word in a sentence, so we can use the tags assigned to preceding words as features.

- At training time, we use the correct IOB tags of preceding terms as features.

- At classification time, we use the predicted IOB tags of preceding terms.

Word	Label
American	B_ORG
Airlines	I_ORG
a	O
unit	O
of	O
AMR	B_ORG
Corp.	I_ORG
immediately	O
matched	O
the	O
move	O
spokesman	O
Tim	B_PERS
Wagner	I_PERS
said	O

IOB Encoder Features

The feature matrix looks something like this:

Word	PoS	Shape	Phrase	…	Prev. Tag	Tag
American	NNP	cap	B_NP		<None>	B_ORG
Airlines	NNPS	cap	I_NP		B_ORG	I_ORG
a	DT	lower	O		I_ORG	O
unit	NN	lower	B_NP		O	O
of	IN	lower	I_NP		O	O
AMR	NNP	upper	B_NP		O	B_ORG

…

Commercial NER

A full production pipeline for NER will combine a few approaches:

1. First, use high-precision rules to tag unambiguous entities
    - e.g. hand-tailored regular expressions
    - Or write parsers for particular web sites, such as Wikipedia
2. Search for substring matches of previously detected names, using probabilistic string-matching metrics
3. Consult application-specific name lists to identify likely name entity mentions from the given domain
4. Apply probabilistic sequence labeling using the tags from 1-3 as additional features

Relation Extraction

Named Entity Recognition | **Relation Extraction**
Question Answering | Summarization

Relation Extraction

- Once we know the entities in a text, we want to know what the text is saying about those entities.
- One part of this is identifying the relationships between the entities.

Citing high fuel prices, **[ORG United Airlines]** said **[TIME Friday]** it has increased fares by **[MONEY $6]** per round trip on flights to some cities also served by lower-cost carriers. **[ORG American Airlines]**, a unit of **[ORG AMR Corp.]**, immediately matched the move, spokesman **[PERS Tim Wagner]** said.

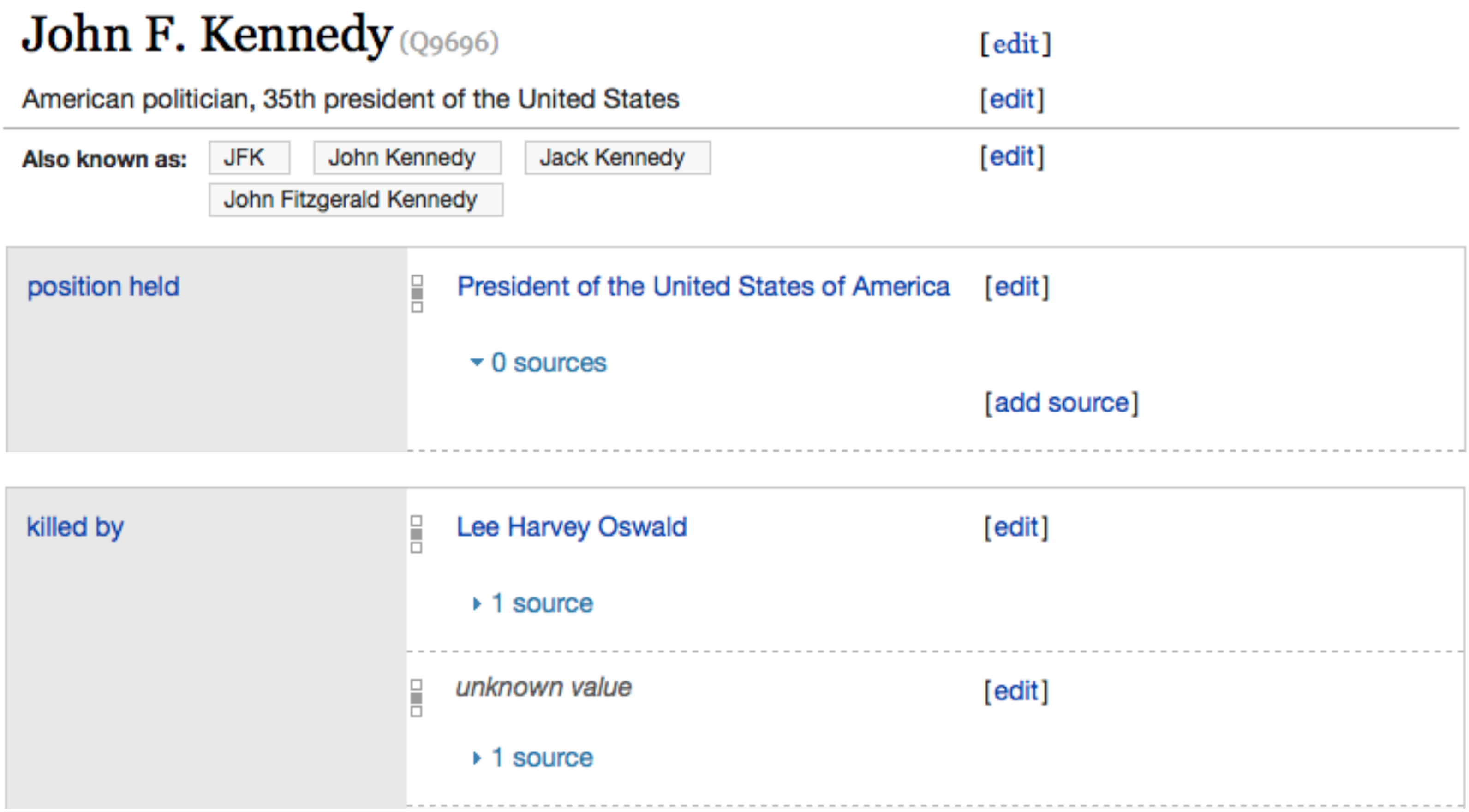
Entity	Relation	Entity
American Airlines	part of	AMR Corp.
Tim Wagner	spokesman for	American Airlines

Knowledge Bases

- There are many existing databases of entity relations, some of which are freely available online:
  - Freebase – Started in 2007, acquired by Google in 2010. Publicly viewable and editable.
  - Knowledge Graph – Google's private version of Freebase, with proprietary data included. Used to populate boxes in Google Search.
  - wikidata – Run by Wikimedia. Publicly editable and viewable.
- However, these databases are also fairly sparse. Many possible relations are missing, and some of their data is unreliable.

wikidata on John F. Kennedy

Some example relations for John F. Kennedy:



John F. Kennedy (Q9696) [edit]

American politician, 35th president of the United States [edit]

Also known as: JFK | John Kennedy | Jack Kennedy | John Fitzgerald Kennedy [edit]

position held	President of the United States of America	[edit]
	0 sources	[add source]
killed by	Lee Harvey Oswald	[edit]
	1 source	
	unknown value	[edit]
	1 source	

Relation Types

Here is a handful of possible relation types.

Relations		Examples	Entity Types
Affiliations	Personal	married to, mother of	PER → PER
	Organizational	spokesman for, president of	PER → ORG
	Artifactual	owns, invented, produces	(PER | ORG) → ART
Geospatial	Proximity	near, on outskirts	LOC → LOC
	Directional	southeast of	LOC → LOC
Part-Of	Organizational	a unit of, parent of	ORG → ORG
	Political	annexed, acquired	GPE → GPE

Extracting Relations

- With types expressed in this way, we can think of a relationship as a tuple: (entity1, relation, entity2).

- Having already identified the entities in a block of text, we now want to identify the relationship tuples implied by the text.

- This can be done as a two step process:

    1. Identify entity pairs which are likely to be related

    2. Determine the relationship type

- Each step involves training a ML classifier. The most important question is: which features should we use?

Features for Relations

- Features of the entities:
    - Entity types, both individually and concatenated (why?)
    - Main words of the entity phrases
    - Bag of words statistics for the entities
- Features of the text around the entities. This is often divided into three groups of features: before the first entity, after the second, and in between the entities.
- Syntactic structure features. Is one in a phrase which modifies the phrase the other was found in? What is the parse-tree distance between the entities?

IR-based Relation Extraction

- An entirely different approach to identifying relations is to define a collection of phrases which identify the relationship type and use an IR system to find occurrences of them.

- For instance, to identify cities in which an airline has a hub, you could search for: “* has a hub at *”

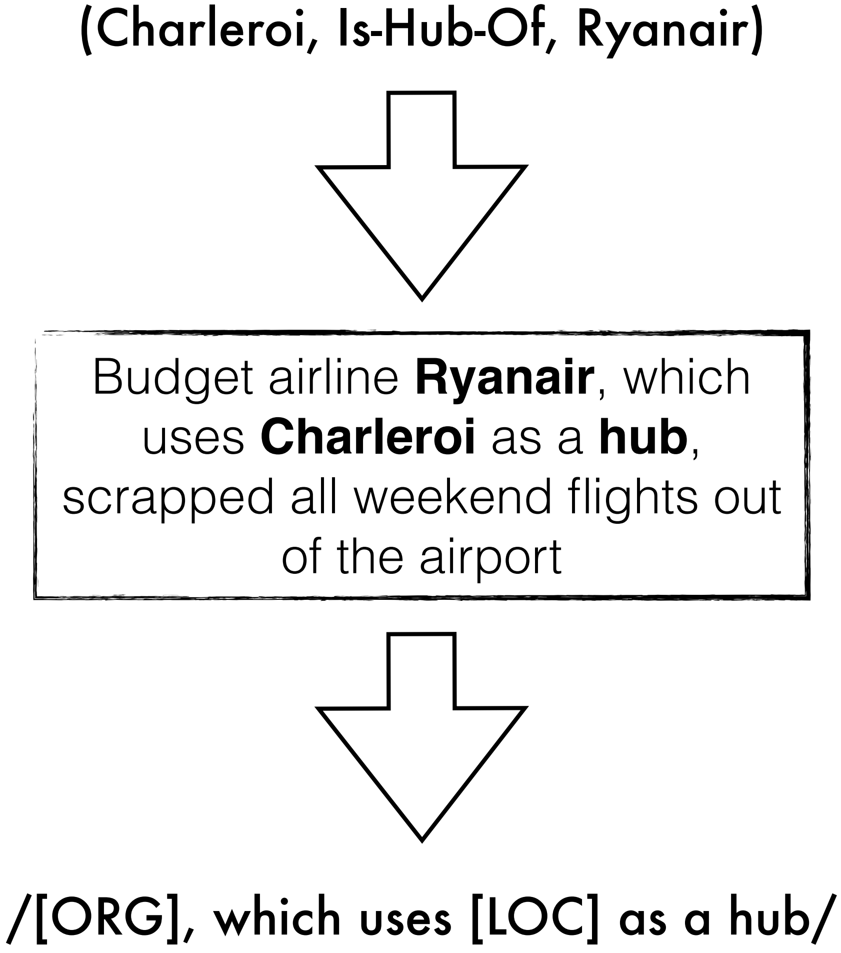
Search Results		
Milwaukee-based Midwest	has a hub at	KCI
Delta	has a hub at	LaGuardia
Bulgaria Air	has a hub at	Sofia Airport, as does Hemus Air
American Airlines	has a hub at	the San Juan airport

IR-based Relation Extraction

- This has some quality problems:
  - FP: “A star topology often has a hub at its center.”
  - FN: “Ryanair also has a continental hub at Charleroi airport (Belgium)”
- False positives can be reduced by filtering using entity type: “[ORG] has a hub at [LOC]”
- False negatives can be reduced by expanding the set of search patterns

Bootstrapping Patterns

- Start with a set of handmade seed patterns
- Run the search, and build a database of relations
- Now search for the known relations in other contexts to discover new effective search patterns
- Repeat as needed



(Charleroi, Is-Hub-Of, Ryanair)

Budget airline **Ryanair**, which uses **Charleroi** as a **hub**, scrapped all weekend flights out of the airport

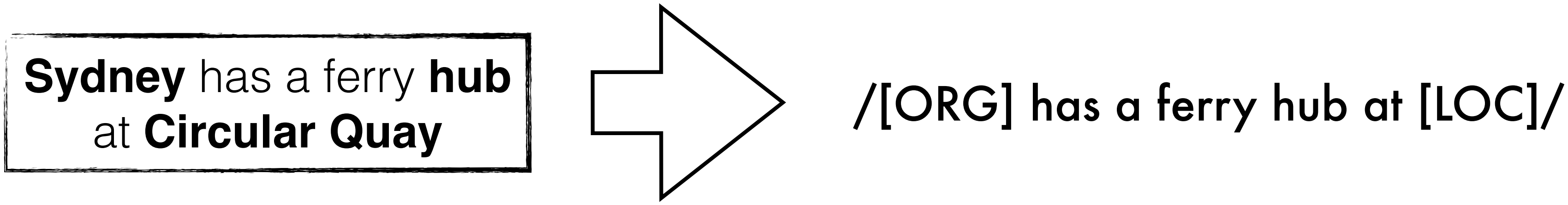
/[ORG], which uses [LOC] as a hub/

Bootstrapping Issues

- How do you extract meaningful patterns from text?

- How do you assess the reliability of patterns, or discovered relations?

- If you’re not careful, bootstrapping can result in **semantic drift**:



Sydney has a ferry **hub** at **Circular Quay**

/[ORG] has a ferry hub at [LOC]/

- This can be mitigated somewhat by carefully comparing the tuples found by different patterns. We expect to see some overlap between the tuples identified by high quality patterns.

Question Answering

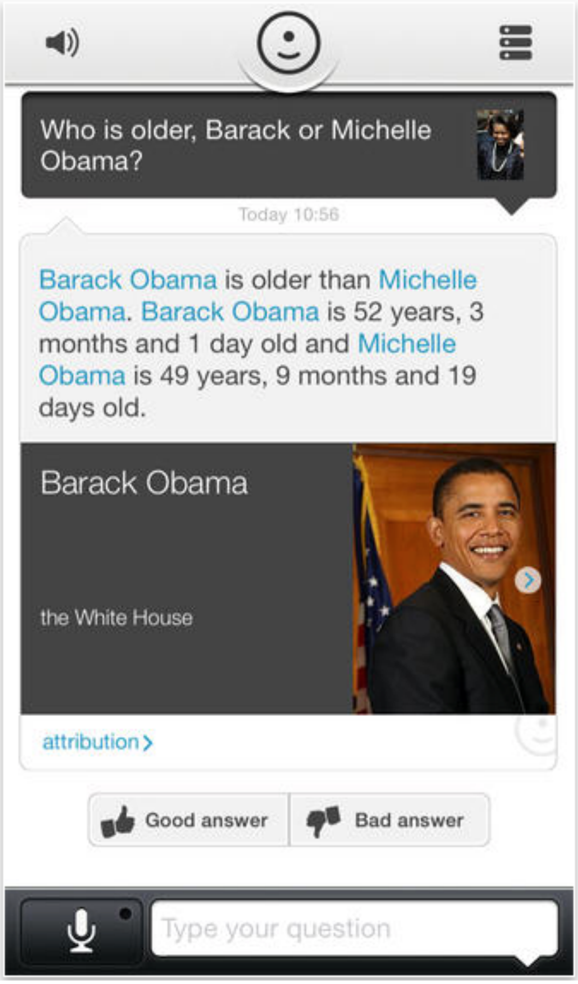
Named Entity Recognition | Relation Extraction
Question Answering | Summarization

Question Answering

- We have mostly focused on **ad-hoc search**, in which a ranked list of documents is presented as a response to a keyword query.

- In **Question Answering**, we instead respond to a direct question with a simple statement of fact.

- This is used in Google search results, and is very popular in mobile apps such as Siri, Evi, and Google Now.



Who is older, Barack or Michelle Obama?

Today 10:56

Barack Obama is older than Michelle Obama. Barack Obama is 52 years, 3 months and 1 day old and Michelle Obama is 49 years, 9 months and 19 days old.

Barack Obama

the White House

Evi Screenshot

Question Answering

The answers can be produced in a number of ways:

- By storing possible answers in a Knowledge Base and converting questions into queries
- By performing logical inference on information stored in a Knowledge Base
- By performing ad-hoc search with the question and post-processing the results
- By delegating certain searches to specialized search engines, such as Wolfram Alpha or IMDB

Factoid Answering with IR

- We will focus on the problem of Factoid Question Answering using IR techniques

- A factoid is a single-sentence statement of fact, usually involving a named entity.

    - e.g. “Barack Obama was born in 1961.”

- Our task is to generate a query for ad-hoc search and then identify a sentence in the top *k* documents which contains the desired factoid.

Querying for Factoids

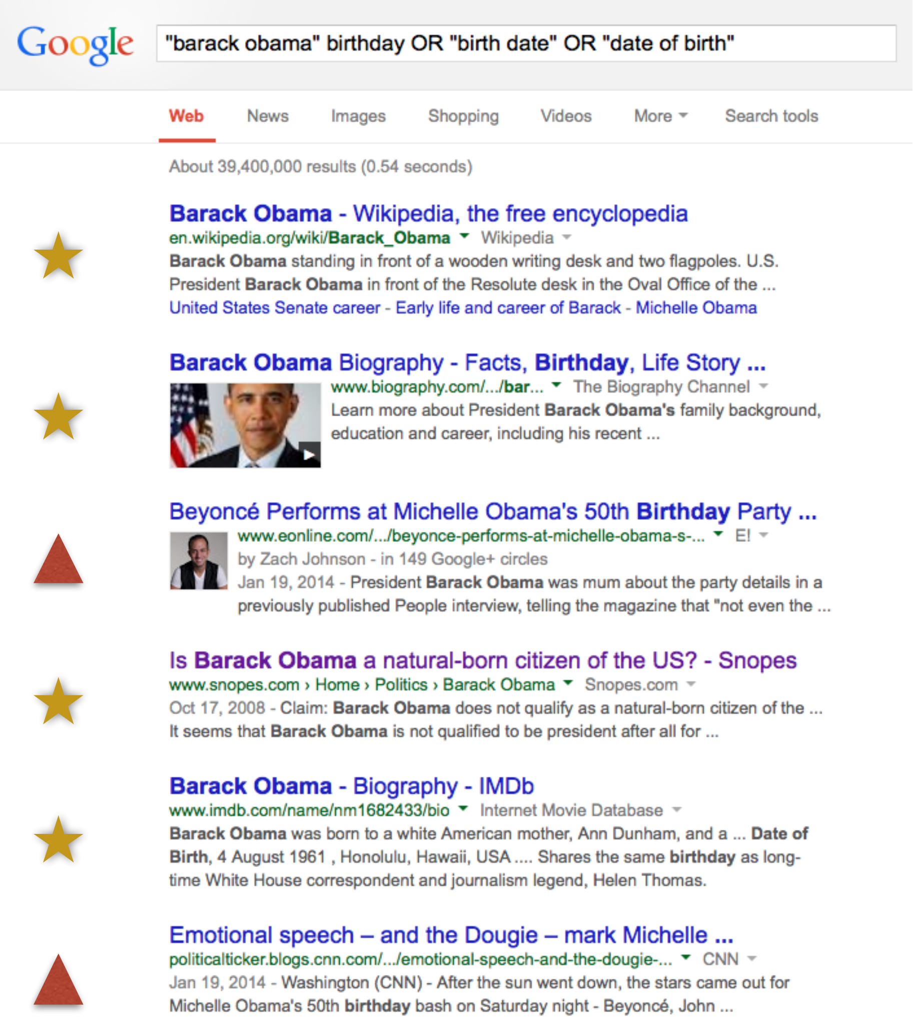
- Ad-hoc search is based on keyword search, and we are starting with a full sentence.

  "What is Obama's birth date?"

- We want to find answers, not questions, so we generally remove the question words ("what")

- We may also want to do query expansion, or even the more expensive NER and entity disambiguation, to improve result quality.

  "Barack Obama" AND ("birth date" OR birthday OR "date of birth")



"barack obama" birthday OR "birth date" OR "date of birth"

About 39,400,000 results (0.54 seconds)

★ **Barack Obama** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Barack_Obama** ▾ Wikipedia ▾
Barack Obama standing in front of a wooden writing desk and two flagpoles. U.S. President **Barack Obama** in front of the Resolute desk in the Oval Office of the ...
United States Senate career - Early life and career of Barack - Michelle Obama

★ **Barack Obama** Biography - Facts, **Birthday**, Life Story ...
www.biography.com/.../**bar**... ▾ The Biography Channel ▾
Learn more about President **Barack Obama's** family background, education and career, including his recent ...

▲ Beyoncé Performs at Michelle Obama's 50th **Birthday** Party ...
www.eonline.com/.../beyonce-performs-at-michelle-obama-s-... ▾ E! ▾
by Zach Johnson - in 149 Google+ circles
Jan 19, 2014 - President **Barack Obama** was mum about the party details in a previously published People interview, telling the magazine that "not even the ...

★ Is **Barack Obama** a natural-born citizen of the US? - Snopes
www.snopes.com › Home › Politics › Barack Obama ▾ Snopes.com ▾
Oct 17, 2008 - Claim: **Barack Obama** does not qualify as a natural-born citizen of the ...
It seems that **Barack Obama** is not qualified to be president after all for ...

★ **Barack Obama** - Biography - IMDb
www.imdb.com/name/nm1682433/bio ▾ Internet Movie Database ▾
Barack Obama was born to a white American mother, Ann Dunham, and a ... **Date of Birth**, 4 August 1961 , Honolulu, Hawaii, USA .... Shares the same **birthday** as long-time White House correspondent and journalism legend, Helen Thomas.

▲ Emotional speech – and the Dougie – mark Michelle ...
politicalticker.blogs.cnn.com/.../emotional-speech-and-the-dougie-... ▾ CNN ▾
Jan 19, 2014 - Washington (CNN) - After the sun went down, the stars came out for Michelle Obama's 50th **birthday** bash on Saturday night - Beyoncé, John ...

Question Classification

- Given a set of candidate documents, the next task we need to perform is identifying the type of answer we expect.

- One approach is to train a classifier to predict the expected entity type of an answer:

  who → PERSON
  when → DATE

- Or to use hand-tailored rules on more complex ontologies:

  What is [PERSON]'s birthday? → PERSON:BIRTH_DATE

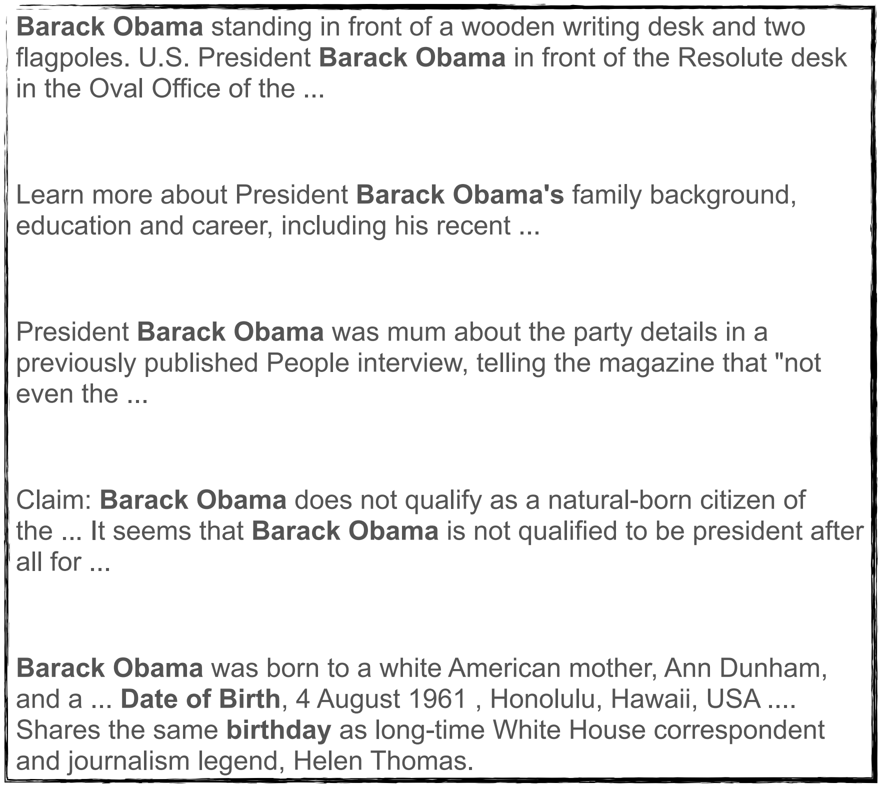
- Most modern systems use supervised learning techniques.

Passage Retrieval

- Now we know what we're looking for (e.g. a relation between [PERSON Barack Obama] and [PERSON:BIRTH-DATE]) and we have a collection of documents to search.

- We want to identify passages in those documents which could serve as suitable answers, rank them, and return the best passage.

- This is similar to the NLP relation extraction and IR snippet generation tasks. In fact, some systems just rank the snippets.



Barack Obama standing in front of a wooden writing desk and two flagpoles. U.S. President **Barack Obama** in front of the Resolute desk in the Oval Office of the ...

Learn more about President **Barack Obama's** family background, education and career, including his recent ...

President **Barack Obama** was mum about the party details in a previously published People interview, telling the magazine that "not even the ...

Claim: **Barack Obama** does not qualify as a natural-born citizen of the ... It seems that **Barack Obama** is not qualified to be president after all for ...

Barack Obama was born to a white American mother, Ann Dunham, and a ... **Date of Birth**, 4 August 1961 , Honolulu, Hawaii, USA .... Shares the same **birthday** as long-time White House correspondent and journalism legend, Helen Thomas.

Snippets of top 5 results

Filtering Passages

- In order to perform passage retrieval, we first need to divide the documents into passages (e.g. paragraphs, sentences, etc.)

- We then filter out passages which are unlikely to contain the answer.

  - Does the passage contain the entity we were asked about and the entity type we're looking for?

- Given a filtered list of passages, we use a machine learning classifier trained to determine passage relevance (e.g. a Learning to Rank classifier)

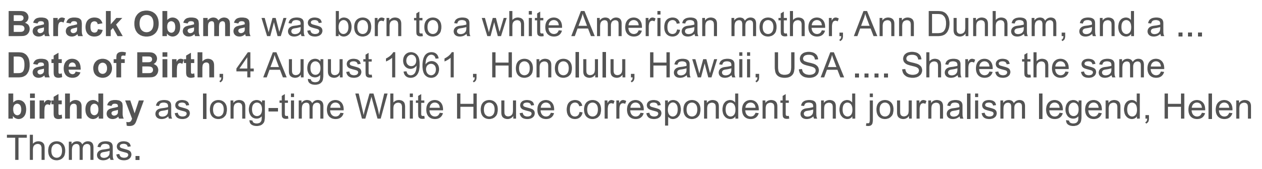
Features for Passages

We might use the following features to identify relevant passages:

- The number of named entities of the desired type in the passage
- The number of question keywords in the passage
- The longest exact sequence of question keywords
- The rank of the document from which the passage was extracted
- The proximity of question keywords to each other
- The ngram overlap between the passage and the question

Answer Processing

- We finally have the top-ranked passage, which we believe contains the answer.



Barack Obama was born to a white American mother, Ann Dunham, and a ... **Date of Birth**, 4 August 1961 , Honolulu, Hawaii, USA .... Shares the same **birthday** as long-time White House correspondent and journalism legend, Helen Thomas.

- We generally want to extract the answer and form our own response.

- With the **pattern extraction** method, we use the expected relation or answer type with hand tailored or bootstrapped patterns to find the answer.

N-gram Tiling Method

The **n-gram tiling** method is a different approach, based on ngram overlap in retrieved passages.

1. First, we extract all unigrams, bigrams, and trigrams from the relevant passages
2. Each n-gram is assigned a weight based on the number of passages in which it occurs
3. N-grams which do not include the expected entity type are filtered out
4. Remaining n-grams which contain overlapping terms are tiled to form a final response
5. The process is repeated with slightly different queries, tiling together high-scoring responses, until some termination point is reached
6. The final answer is shown to the user

Summarization

Named Entity Recognition | Relation Extraction
Question Answering | **Summarization**

Summarization

- Factoid question answering provides bite sized answers to simple factual questions, but some questions are more complex.

- In **text summarization**, we distill the most important information from a text to present for a particular task and user.

- Current research focuses on a few types, listed to the right.

Summarization Types
outlines of any document
abstracts of a scientific article
headlines of a news article
snippets summarizing a Web page on a search results list
action items or other summaries of a business meeting
summaries of e-mail threads
compressed sentences for producing simplified text
answers to complex questions, summarizing multiple documents

Summarization Taxonomy

- Summaries may be of a **single** document, or **multiple** documents.

- An **extract** presents a subset of text from the summarized documents, while an **abstract** generates new text to describe them.

- A **generic summary** focuses on giving the important information in the documents, without respect to a particular user or information need.

- **Query-focused summarization** instead tries to present information relevant to some user or query.

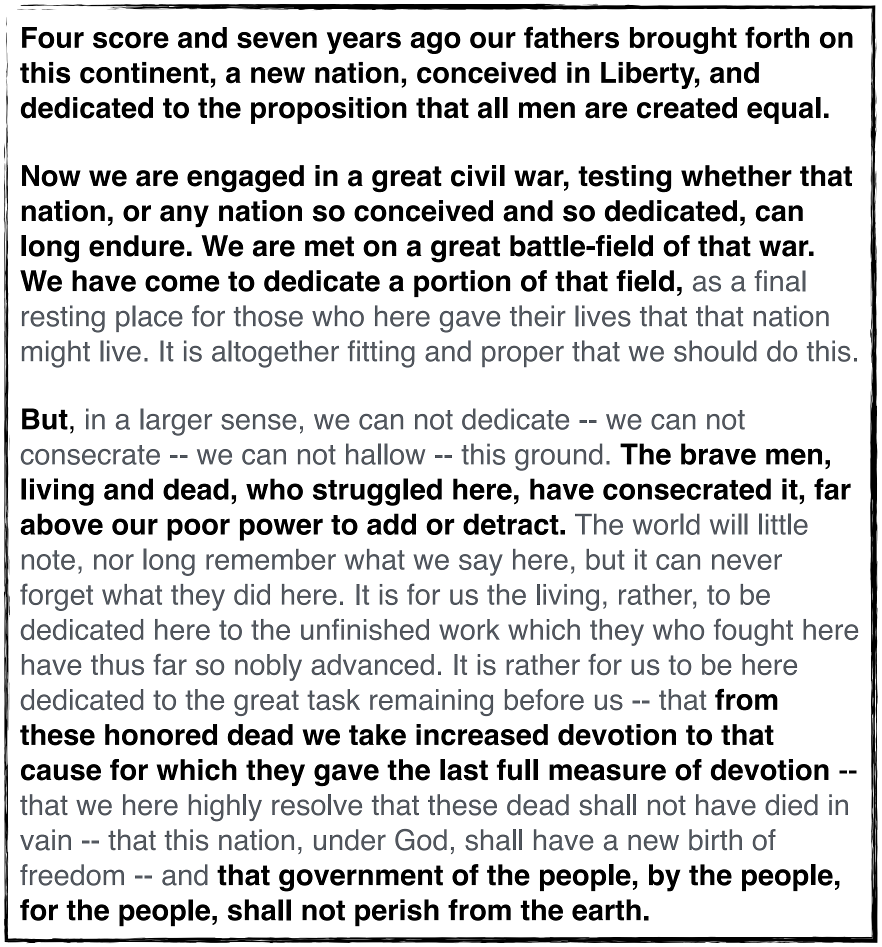
Single Document Extracts

To generate an extract summarizing a single document, we need to carry out several steps.

1. **Content Selection:** Which sentences should we include?

2. **Information Ordering:** How should we order and structure the sentences?

3. **Sentence Realization:** Final cleanup to produce a fluent summary.



Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we can not dedicate -- we can not consecrate -- we can not hallow -- this ground. **The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract.** The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us -- that **from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion --** that we here highly resolve that these dead shall not have died in vain -- that this nation, under God, shall have a new birth of freedom -- and **that government of the people, by the people, for the people, shall not perish from the earth.**

Content Selection in Gettysburg Address

Unsupervised Selection

- A simple unsupervised approach to content selection is to choose the top *k* sentences, ranked by some weight function.
    - Average cosine similarity to other sentences in the document
    - Average TF-IDF of the words in a sentence
    - Average log-likelihood ratio of the words in a sentence, using a threshold to approximate statistical significance:

$$weight(s_i) = \frac{\sum_{w \in s_i} llrw(w)}{|\{w : w \in s_i\}|}$$

$$llrw(w) = \begin{cases} 1 & \text{if } -2\log(llr(w)) > 10 \\ 0 & \text{otherwise} \end{cases}$$

Supervised Selection

For a supervised approach to content selection, we need humans to tag sentences for inclusion in an extract. We then train a binary classifier to combine features for selection.

Feature	Description
position	The position of the sentence in the document. For instance, the first and last sentences are often good choices.
cue phrases	Presence of phrases like *in summary*, *in conclusion*, or *it seems to me that*
word informativeness	Indicators of topical relevance, such as query relevance or presence in topic signature
sentence length	Whether the sentence is too short to be useful
cohesion	Whether the sentence includes many words distinctive to the document

Sentence Simplification

- The sentences we select will often contain irrelevant information, and a good summary will shorten them.

- One approach is to parse the sentence and apply rules to remove certain structures.

Structure	Example
appositives	Rajam, ~~28, an artist who was living at the time in Philadelphia,~~ found the inspiration in a magazine.
attribution clauses	Rebels agreed to talks on Tuesday, ~~according to government officials.~~
PPs without named entities	The commercial fishing restrictions will not be lifted unless the salmon population increases ~~to a sustainable number.~~
initial adverbials	“For example,” “On the other hand,” “As a matter of fact”

Multiple Documents

- The biggest problem introduced by summarizing multiple documents is avoiding redundant sentences.
- One approach is to penalize sentences at selection time if they are very similar to previously-selected sentences.
- Alternatively, you could cluster sentences based on their content and select a centroid sentence from each cluster.
- You could also use clusters to assist in simplification, by taking advantage of different phrasings of redundant sentences.

Information Ordering

- In single-document summarization, we can generally use the sentence order from the original document.

- Ordering sentences coherently from a multi-document summarization is a fairly difficult problem. Many sentence permutations are confusing, or even misleading.

- Finding a good sentence ordering involves some model of **text coherence**, which is itself a complex NLP subject.

Text Coherence

- Readers try to form connections between successive sentences, and get confused when this process fails.
- These may be causal:
  - John hid Bill's car keys. He was drunk.
  - John hid Bill's car keys. He likes spinach.
- They may be drawing analogy:
  - The Scarecrow wanted some brains. The Tin Woodsman wanted a heart.
- They may be narrative:
  - Dorothy picked up the oil-can. She oiled the Tin Woodsman's joints.

Toward Coherence

- Several strategies have been developed which help produce more coherent orderings.
  - Prefer orderings that result in sensible coherence relations between sentences.
  - Prefer orderings that place sentences with similar important terms near each other.
  - Prefer ordering with an orderly transition between named entities.
  - Use a template for question answering that presents sentences in a human-designed order.

Focused Summarization

- **Focused Summarization** is generating a summary in response to a query

- A straightforward approach is to identify relevant documents, then include sentence-level query relevance as a feature for sentence selection

- Another approach is to pre-define certain categories of information needs, and create templates for their answers. This might employ a knowledge base as an information source.

Focused Summarization

- **Focused Summarization** is generating a summary in response to a query

- A straightforward approach is to identify relevant documents, then include sentence-level query relevance as a feature for sentence selection

- Another approach is to pre-define certain categories of information needs, and create templates for their answers. This might employ a knowledge base as an information source.

Category	Example
genus	The Hajj is a type of ritual
species	the annual hajj begins in the twelfth month of the Islamic year
synonym	The Hajj, or Pilgrimage to Mecca, is the central duty of Islam
subtype	Qiran, Tamattu', and Ifrad are three different types of Hajj

Definition Template

Summary

- Finding particular information is much more difficult than ad hoc search (for most ad hoc queries).

- A common theme in all these tasks is finding effective features for classification tasks.

- Many of these tasks were initially developed using hand-tailored rules, which were later replaced with more data-driven machine learning approaches.

- There is plenty of room for improvement in these techniques. All of them are actively researched.