Information extraction

Vivi Nastase

Summer semester 2012, ICL, University of Heidelberg

What is information extraction?

- what is "information"?
- why do we want it?
- how do we extract it?

What is “information” and why do we want it?

What is "information" and why do we want it?

Knowledge obtained from investigation, study or instruction (M-W), in our case:

- "things" we talk about
- Attributes of "things"
- Relations between "things"

What is "information" and why do we want it?

Democratic Senator Barack Obama has been elected the first black president of the United States, prompting celebrations across the country.
"It's been a long time coming, but tonight... change has come to America," the president-elect told a jubilant crowd at a victory rally in Chicago.
His rival John McCain accepted defeat, and called on his supporters to lend the next president their goodwill.

What is "information" and why do we want it?

Senator Barack Obama president United States celebrations country. time change America president-elect crowd victory rally Chicago. rival John McCain defeat supporters president goodwill.

What is "information" and why do we want it?

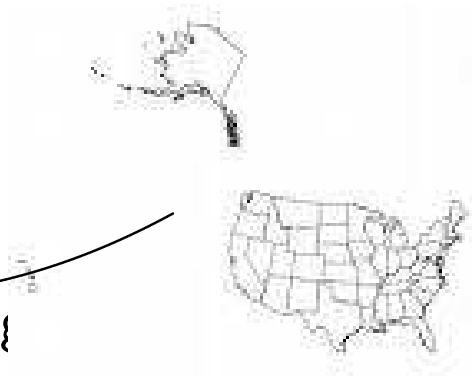
Senator Barack Obama president United States celebrations country.

time change America president-elect crowd victory rally Chicago.

rival John McCain defeat supporters president goodwill.

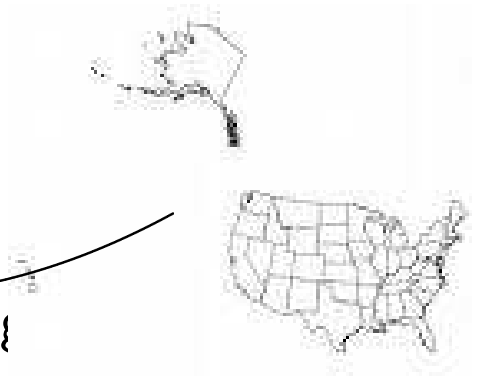
What is "information" and why do we want it?

Senator Barack Obama
president United States
country.
time
chan
America president-elect crowd
victory rally Chicago.
rival John McCain defeat sup-
porters president goodwill.

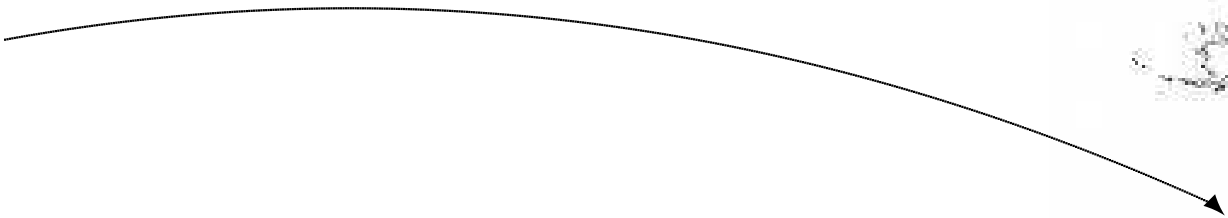


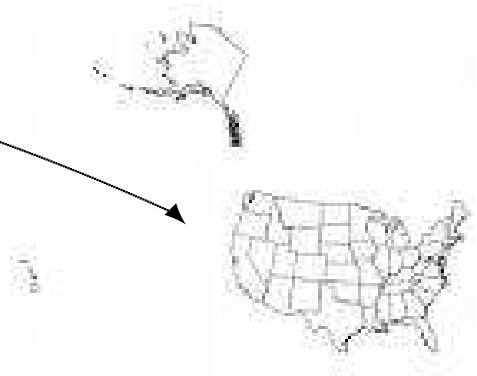
What is "information" and why do we want it?

Senator Barack Obama
president United States
country.
time chan
America president-elect crowd
victory rally **Chicago**.
rival **John McCain** defeat sup-
porters president goodwill.



What is "information" and why do we want it?





Barack Obama

United States

Barack H. Obama **is** the 44th President of the United States. (whitehouse.org)

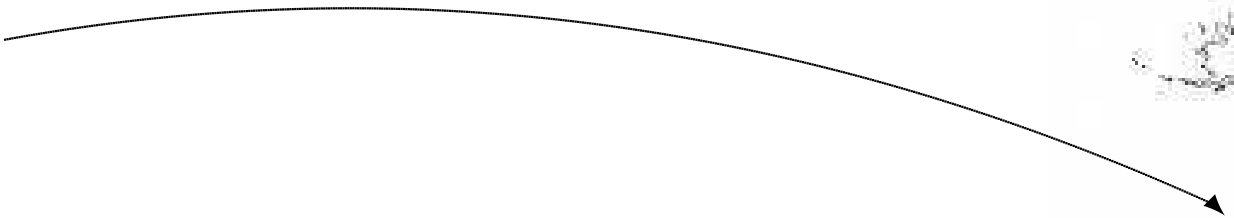
Now, therefore, I, Barack Obama, President of the United States, ... (whitehouse.org)

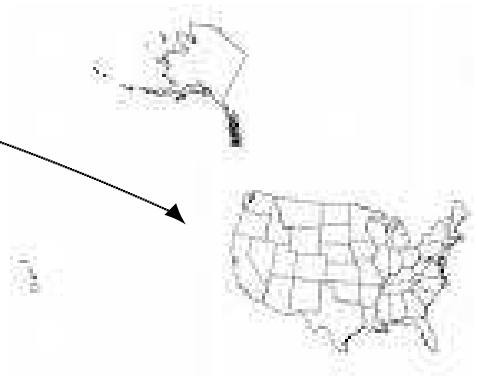
Barack Obama **Is Elected** 44th President Of The United States Of America. (youtube.com)

Barack Obama (president of United States), August 4, 1961 Honolulu, Hawaii (britannica.com)

...

What is "information" and why do we want it?





Barack Obama

United States

French President Nicolas Sarkozy faces an uphill struggle in the second round of the presidential election

Nicolas Sarkozy (president of France), Jan. 28, 1955 Paris ...

In 2011 Dilma Rousseff **became** Brazil's first woman president ...

Johnston will meet Brazilian President Dilma Rousseff, attend meetings ...

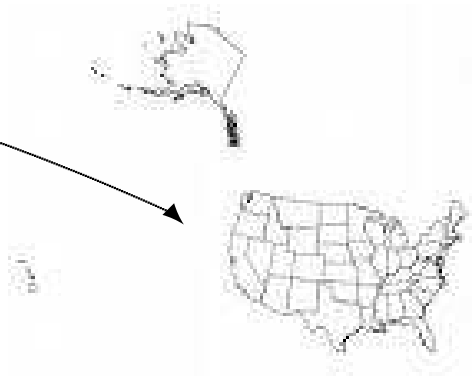
Vclav Klaus **is** the second President of the Czech Republic ...

...

What is "information" and why do we want it?

president
president-elect

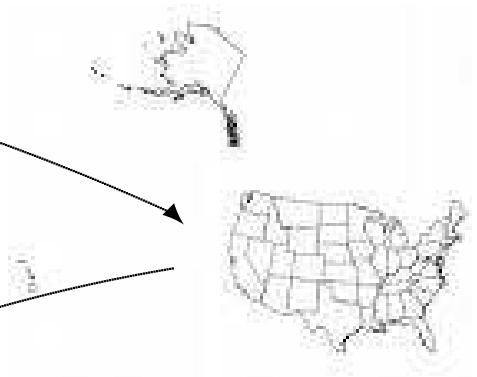
Barack Obama



United States

What is "information" and why do we want it?

president
president-elect



Barack Obama

located in

United States
America

senator

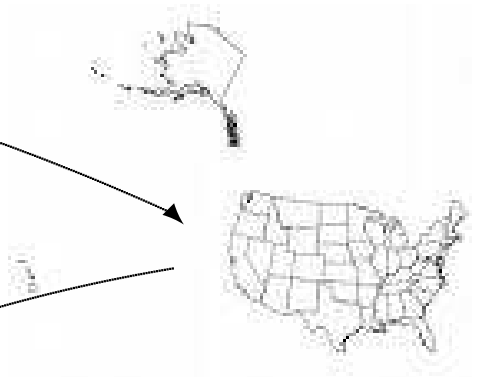
is a

Chicago

country

What is "information" and why do we want it?

president
president-elect



Barack Obama

located in

United States
America

senator

is a

tonight?

Chicago

country

"Things": instances and concepts

instance/entity = objects in the world
Barack Obama, *John McCain*, *USA*, *America*, ...

concept/class = a placeholder for a set of instances (objects) that share similar properties. They may have definitions and names (but not necessarily!)
president, *senator*, *country*, ...

"Things": instances and concepts and events?

instance/entity = objects in the world
Barack Obama, John McCain, USA, America, ...

concept/class = a placeholder for a set of instances (objects) that share similar properties. They may have definitions and names (but not necessarily!)
president, senator, country, ...

events = "things" that happen, have causes and consequences, time frame, participants, attributes
elected, told, ...
celebration, defeat, ...

Relations between "things"

Relations are assertions linking n concepts:
person president of country; cities capital of countries

Facts are instantiations of relations, linking instances of concepts:
Barack Obama president of USA; Berlin capital of Germany

Attributes correspond to facts that capture quantifiable properties of a class or instance:
car $\rightarrow$ weight, max speed, consumption
election $\rightarrow$ date, result, ...

How can we obtain this information?

"things"

- terminology analysis
- keyphrase extraction
- named entity (NE) recognition and disambiguation

relations

attributes

How can we obtain this information?

"things"

relations

- bootstrapping using patterns and seed examples
- detection and classification

attributes

How is this information used?

"things"	• build gazeteers, lexica, ...
relations	• mining for specific relations, for specific tasks (e.g. template filling – Who? What? Where? When?) • build ontologies • knowledge acquisition ...
attributes	• query processing (e.g. class attributes) • knowledge acquisition ...

Information retrieval, information extraction, question answering?

Course plan

Scheduling:

- Lecture: Mondays, 14-16, here
- Office hours: Thursdays, 11-12 (Room 121)
- e-mail: nastase@cl.uni-heidelberg.de

Work:

- attend the lectures, and interact
- a semester long project – evaluation competition at the end
- presenting and discussing an assigned paper
- oral exam

Project: Temporal expression analysis and timeline construction

July 11, 2011
NASA's Proud Space Shuttle Program Ends With Atlantis Landing
Agency Ushers in Next Era of Exploration

CAPE CANAVERAL, Fla. – Wrapping up 30 years of unmatched achievements and blazing a trail for the next era of U.S. human spaceflight, NASA's storied Space Shuttle Program came to a "wheels stop" on Thursday at the conclusion of its 135th mission.

Shuttle Atlantis and its four-astronaut crew glided home for the final time, ending a 13-day journey of more than five million miles with a landing at 5:57 a.m. EDT at NASA's Kennedy Space Center in Florida. It was the 26th night landing (20th night and 78th total landings at Kennedy) and the 133rd landing in shuttle history.

...

Temporal expression analysis

July 11, 2011
NASA's Proud Space Shuttle Program **Ends** With Atlantis Landing Agency Ushers in **Next Era** of Exploration

CAPE CANAVERAL, Fla. – **Wrapping up 30 years** of unmatched achievements and blazing a trail for the **next era** of U.S. human spaceflight, NASA's storied Space Shuttle Program came to a "wheels **stop**" on **Thursday** at the **conclusion** of its 135th mission.

Shuttle Atlantis and its four-astronaut crew glided home for **the final time**, **ending** a **13-day** journey of more than five million miles with a landing at **5:57 a.m. EDT** at NASA's Kennedy Space Center in Florida. It was the 26th **night** landing (20th **night** and 78th total landings at Kennedy) and the 133rd landing in shuttle history.

...

Temporal expression analysis

11.07.2011
NASA's Proud Space Shuttle Program Ends With Atlantis Landing
Agency Ushers in Next Era of Exploration

CAPE CANAVERAL, Fla. – **Wrapping up 30 years** of unmatched achievements and blazing a trail for the **next era** of U.S. human spaceflight, NASA's storied Space Shuttle Program came to a "wheels **stop**" on **11.07.2011** at the **conclusion** of its 135th mission.

Shuttle Atlantis and its four-astronaut crew glided home for **the final time**, **ending** a **13-day** journey of more than five million miles with a landing at **11.07.2011.5:57.EDT** at NASA's Kennedy Space Center in Florida. It was the 26th **night** landing (20th **night** and 78th total landings at Kennedy) and the 133rd landing in shuttle history.

...

Timeline construction