YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia (J. Hoffart et al, 2012)

presented by Gabe Radovsky

YAGO2: overview

- temporally/spatially anchored knowledge base
- built automatically from Wikipedia, GeoNames, and WordNet
- contains 447 million facts about 9.8 million entities
- human evaluators judged 97.8% of facts correct

The (original) YAGO knowledge base

- introduced in 2007
- automatically constructed from Wikipedia
  - each article in Wikipedia became an entity
- about 100 manually defined relations
  - e.g. wasBornOnDate, locatedIn, hasPopulation
- used SPO (subject, predicate, object) triples to represent facts
  - reification: every fact given an identifier, e.g. wasFoundIn(fact, Wikipedia)

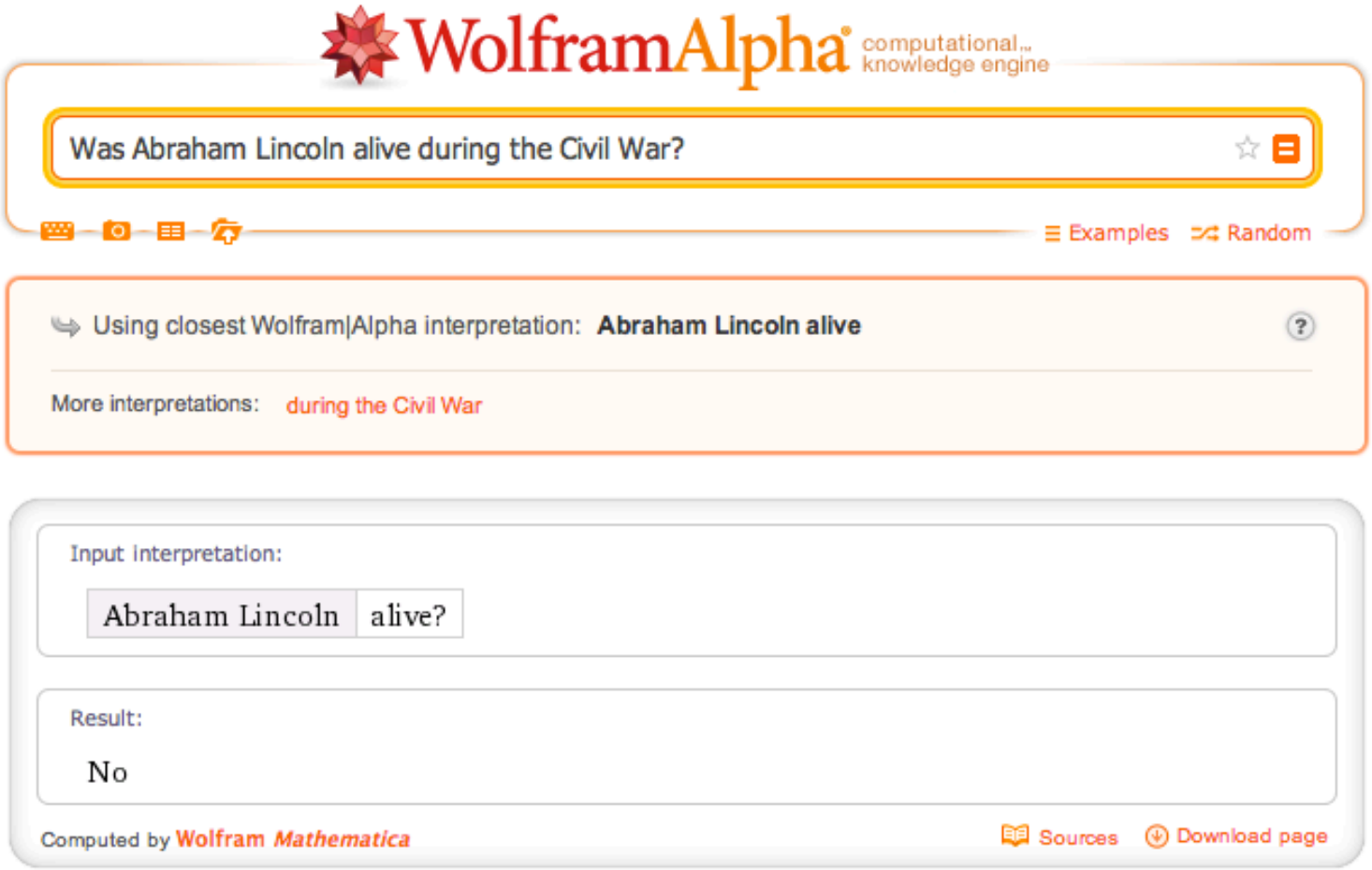
YAGO2: motivation

- WordNet/other lexical resources:
  - manually compiled
  - knows that ‘musician’ is a hyponym of ‘human’; doesn’t know that Leonard Cohen is a musician
- Wikipedia/GeoNames
  - very large collections of (semi-)structured data
  - advances in information extraction make them easier to mine

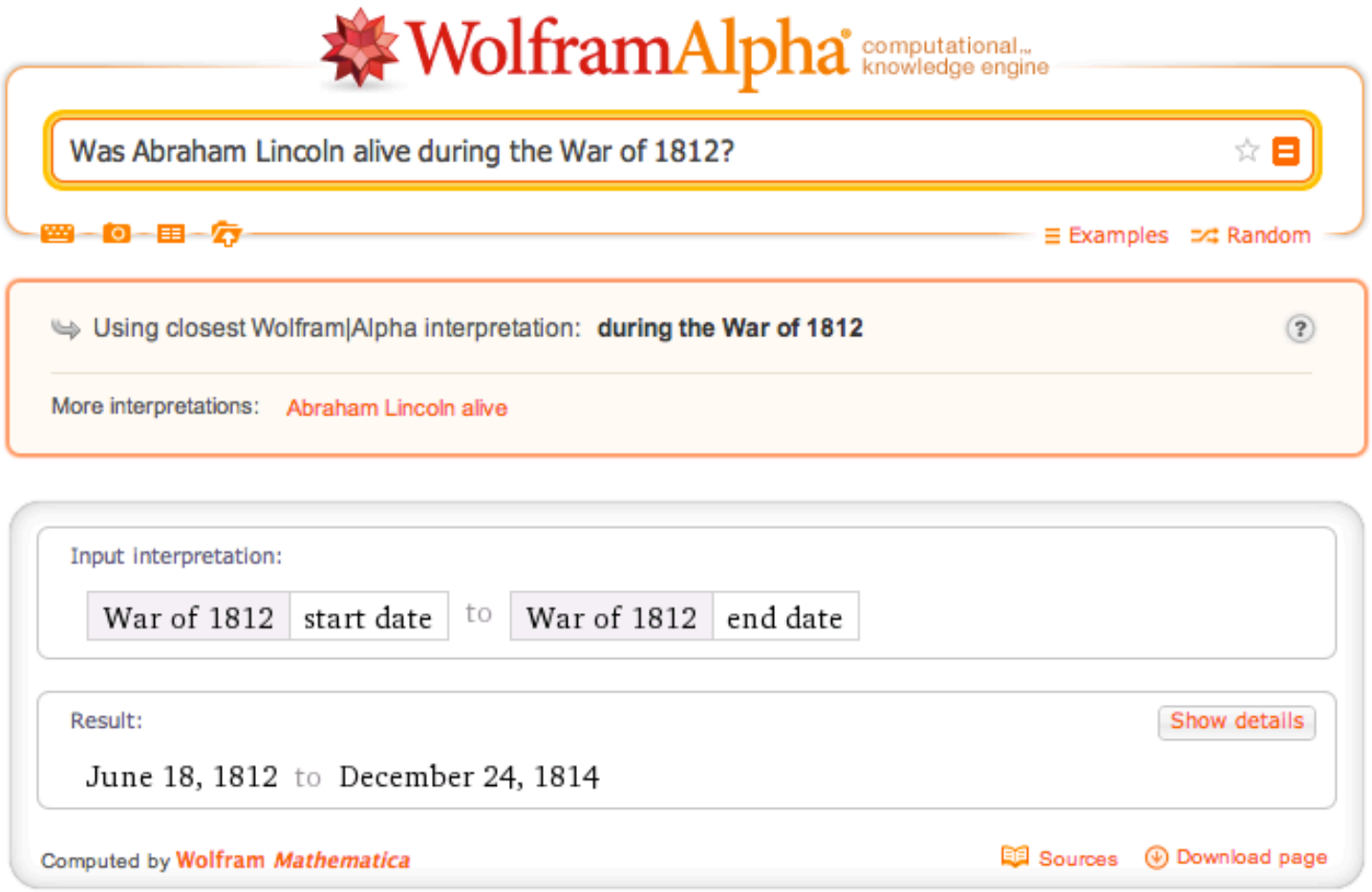
YAGO2: motivation

- “...current state-of-the-art knowledge bases are mostly blind to the temporal dimension” (29).
- e.g. (knowing that Abraham Lincoln was born in 1809 and died in 1865) != (knowing that Abraham Lincoln was alive in 1850)

www.wolframalpha.com



www.wolframalpha.com



YAGO2: contribution

- top-down ontology “with the goal of integrating entity-relationship-oriented facts with the spatial and temporal dimensions” (29).
- new representation model: SPOTL tuples
  - (SPO [subject, predicate, object] + time + location)
- frameworks for extracting knowledge from structured or unstructured text

Extraction architecture for YAGO2

- factual rules
  - "declarative translations of all the manually defined exceptions and facts that the previous YAGO code contained" (30)
- implication rules
  - e.g. if *relation b* is a sub-property of *relation a*, all instances of *b* are also instances of *a*
- replacement rules
  - |"\{\{USA\}\}" replace "[[United States]]"
  - eliminate Wikipedia administrative categories, e.g. "Articles to be cleaned up"

Giving YAGO a temporal dimension

- YYYY-MM-DD format for dates
  - YYYY-##-## if only year is known
- entities
  - given a time span
- facts
  - time point for instantaneous events, time span for events with extended duration
- not all entities/facts could be temporally annotated

Entities and time

- people
  - wasBornOnDate, diedOnDate
- groups, artifacts
  - wasCreatedOnDate, wasDestroyedOnDate
  - some have unbounded end points, e.g. pieces of music, scientific theories
- events
  - startedOnDate, endedOnDate, happenedOnDate (for punctual events)
- entities w/o defined start or end point
  - e.g. numbers, mythological figures, virus strains
  - not assigned temporal information

Facts and time

- facts with an extracted time
  - ElvisPresley diedOnDate 1977-08-16
- facts with a deduced time
  - ([ElvisPresley diedIn Memphis] 1977-08-16)
- extraction time of facts is also included
  - e.g. extractedFrom Wikipedia on YYYY-MM-DD

Giving YAGO a spatial dimension

- YAGO2 “concerned with entities that have a permanent spatial extent on Earth” (34)
  - e.g. countries, cities, mountains, rivers
  - original YAGO, WordNet have no geographical super-class
- new class: yagoGeoEntity
  - type yagoGeoCoordinates stores latitude/longitude pair
- only coordinates, no polygons
  - city center, not exhaustive boundaries

Harvesting geo-entities

- harvested from Wikipedia and GeoNames
- assigned only one class
  - Berlin = “capital of a political entity”
- hierarchical
  - Berlin is located in Germany is located in Europe

Assigning a location

- given to both entities and facts when "ontologically reasonable" (36)
- locations are themselves geo-entities

Entities and location

- events
  - if specific location, e.g. battles and sports competitions
  - happenedIn relation
- groups
  - company headquarters, university campus
  - isLocatedIn relation
- artifacts
  - Mona Lisa in the Louvre
  - isLocatedIn relation

(Con-)textual data in YAGO2

- non-ontological information from Wikipedia (take strings as arguments)
  - hasWikipediaAnchorText (visible text in hyperlink)
  - hasWikipediaCategory
  - hasCitationTitle (from references list)
- multilingual information
  - extracted from inter-language links in articles
  - e.g. [BattleAtWaterloo isCalled SchlachtBeiWaterloo] with associated fact [inLanguage German]

YAGO2: evaluation

- formed one pool for each relation
  - e.g. wasBornOnDate, hasGDP
  - randomly selected test data from each pool
- used 26 human judges
  - judge presented with fact, along with original Wikipedia article to assess its accuracy
    - accuracy of Wikipedia not assessed
  - continued evaluating each pool until confidence interval was smaller than ±5%, to assure statistical significance
- 97.8% of facts were judged correct

YAGO2: evaluation

Table 3
Evaluation of best and worst relations.

Relation	#Total facts	#Evaluated	Accuracy
created	225 563	94	98.04% ± 1.96%
diedIn	28 834	88	97.91% ± 2.09%
happenedOnDate	27 563	94	97.86% ± 2.14%
	⋮		
hasHeight	26 477	120	91.99% ± 4.59%
hasBudget	547	95	90.97% ± 5.41%
hasGDP	175	93	90.79% ± 5.52%

p. 42

Task-based evaluation: Jeopardy

In June 1876 George Custer made his last stand at the Battle of this river.

```
?x isa battle overlaps 1876-06 matches (+George +Custer) .
?x happendIn ?r .
?r isa river
```

It returns the correct result `Battle of the Little Bighorn`.

p. 46

Task-based evaluation: Jeopardy

Q: Disneyland opens & the peace symbol is created
A: 1950s

```
PeaceSymbol wasCreatedOnDate ?x . Disneyland wasCreatedOnDate ?y
```

Result: Correct.

Q: The Empire State Building opens & the "War of the Worlds" radio broadcast causes a panic
A: 1930s

```
EmpireStateBuilding wasCreatedOnDate ?x
```

Result: Correct.

Q: Klaus Barbie is sentenced to life in prison & DNA is first used to convict a criminal
A: 1980s

NA

Result: Not expressible.

Q: The first flight takes place at Kitty Hawk & baseball's first World Series is played
A: 1900s

```
BaseballWorldSeries wasCreatedOnDate ?x
```

Result: Correct.

Q: The first modern crossword puzzle is published & Oreo cookies are introduced
A: 1910s

```
Oreo wasCreatedOnDate ?x
```

Result: Could not be answered, as `Oreo` is not in YAGO2.

Task-based evaluation: Jeopardy

Q: This famed aviator outlived his brother by 35 years, passing away in 1948 on the same day Gandhi was assassinated
A: Orville Wright

```
Gandhi diedOnDate ?d .
?p diedOnDate ?d .
?p type aviator
```

Result: Nearly correct; the YAGO2 result is `WrightBrothers` instead of Orville Wright.

p. 56

Our project

- were originally planning to attempt hierarchical ontology based on Wikipedia
- new project: hierarchical classification of social science journal articles
- mine text of articles with Python NLTK
  - plain text ngrams for n=(1-5)
  - stemmed/POS tagged unigrams
  - possibly named entities
- run different clustering algorithms on entities (article titles with features mined from text)
- attempt to automatically generate reasonable names for clusters