

Information Extraction

Salar Mohtaj | DFKI

Information extraction

- What is information extraction
- Named entity recognition
- Named entity recognition approaches
- NER evaluation metrics

Information extraction

- What is information extraction
- Named entity recognition
- Named entity recognition approaches
- NER evaluation metrics

What is information extraction

- Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured text
- It's the task of finding and understanding limited relevant parts of texts
- Gather information from many pieces of text
- Produce a structured representation of relevant information
- Relations (in the database sense)

What is information extraction

- Goals
 - Clear factual information which is helpful for
 - Answer questions
 - Analytics
- Organize and present information
 - Info boxes in Wikipedia

What is information extraction

roots book

All

Images

Shopping

Videos

Books

More

Settings

Tools

About 764,000,000 results (0,88 seconds)

https://en.wikipedia.org › wiki › Roots:_The_Saga_of_...

Roots: The Saga of an American Family - Wikipedia

Roots: The Saga of an American Family is a 1976 novel written by Alex Haley. It tells the story of Kunta Kinte, an 18th-century African, captured as an adolescent, sold into slavery in Africa, transported to North America; following his life and the lives of his descendants in the United States down to Haley.

LC Class: E185.97.H24 A33

Publisher: Doubleday

Author: Alex Haley

Publication date: August 17, 1976

Plot · Characters in Roots · Reception · Historical accuracy

People also ask

What happened to Alex Haley?

Was Alex Haley related to Kunta Kinte?

Is roots a true story?

How many sons did Chicken George have?

Feedback

https://www.amazon.com › Roots-American-Family-Ale...

Roots: The Saga of an American Family: Haley, Alex ...

Based off of the bestselling author's family history, this novel tells the story of Kunta Kinte, who is sold into slavery in the United States where he and his ...

https://www.goodreads.com › book › show

Roots: The Saga of an American Family by Alex Haley

Roots: The Saga of an American Family is a novel written by Alex Haley and first published in 1976. Roots tells the story of Kunta Kinte—a young man taken from the Gambia when he was seventeen and sold as a slave—and seven generations of his descendants in the United States.

★★★★★ Rating: 4,4 · 150,712 votes

Roots: The Saga of an American Family

Novel by Alex Haley

4,4/5 · Goodreads

93% liked this book

Google users

Roots: The Saga of an American Family is a 1976 novel written by Alex Haley. It tells the story of Kunta Kinte, an 18th-century African, captured as an adolescent, sold into slavery in Africa, transported to North America; following his life and the lives of his descendants in the United States down to Haley. [Wikipedia](#)

Originally published: August 17, 1976

Author: Alex Haley

Pages: 704 pp (First edition, hardback)

Awards: Pulitzer Prize Special Citations and Awards

Adaptations: Roots (1977), Roots (2016), Roots: The Next Generations (1979)

Genres: Novel, Biography, Historical Fiction, Fictional Autobiography

Book quotes

Characters

Rate and review

Roots: The Saga of an American Family is a 1976 novel written by Alex Haley. It tells the story of Kunta Kinte, an 18th-century African, captured as an adolescent, sold into slavery in Africa, transported to North America; following his life and the lives of his descendants in the United States down to Haley.

[Wikipedia](#)

Originally published: August 17, 1976

Author: Alex Haley

Pages: 704 pp (First edition, hardback)

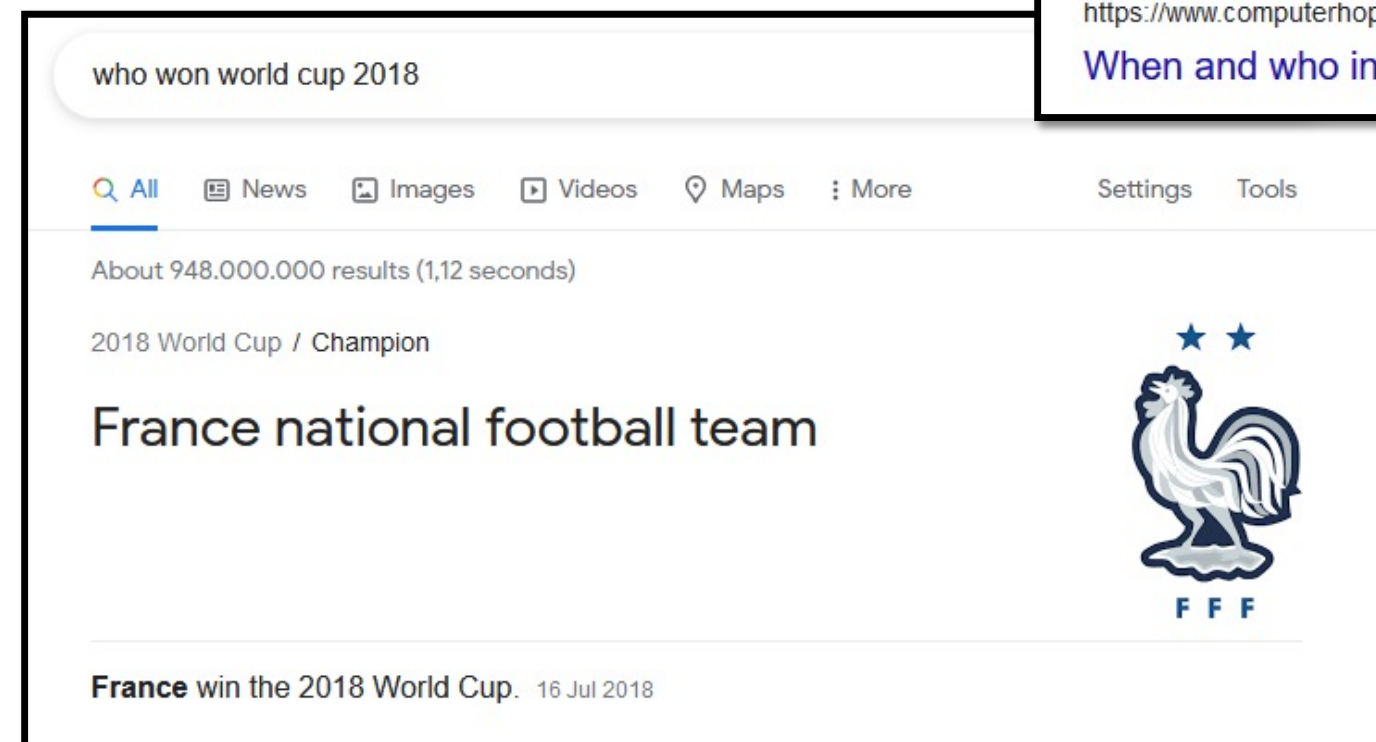
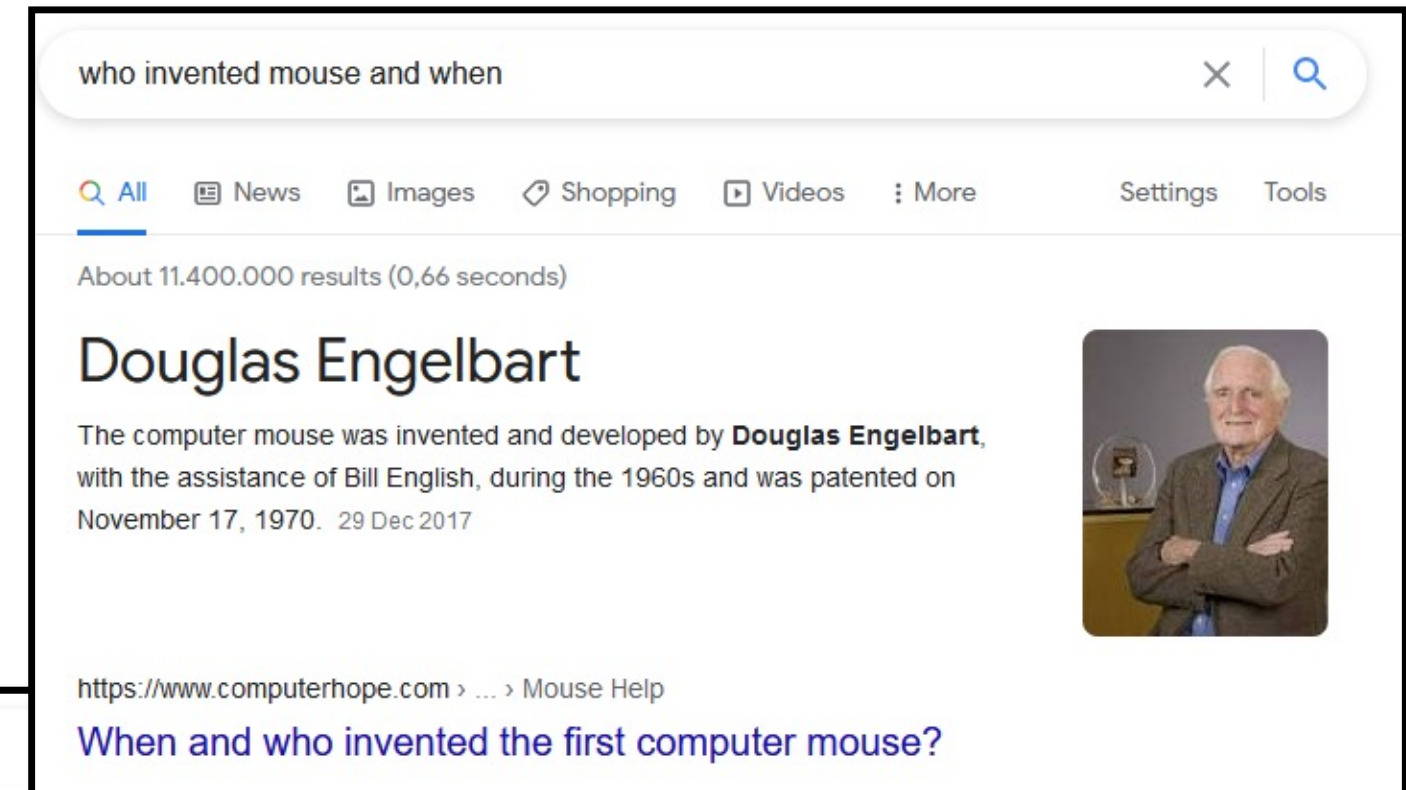
Awards: Pulitzer Prize Special Citations and Awards

Adaptations: Roots (1977), Roots (2016), Roots: The Next Generations (1979)

Genres: Novel, Biography, Historical Fiction, Fictional Autobiography

What is information extraction

- IE systems extract clear, factual information
- Who did what with whom and when?
- Who invented mouse and when
- Who won world cup 2018?



Information extraction

- What is information extraction
- **Named entity recognition**
- Named entity recognition approaches
- NER evaluation metrics

Named entity recognition

- Named Entity Recognition (NER) is a very important sub-task in information extraction: find and classify names in text
- NER seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories
- E.g., person names, organizations, locations, time expressions, quantities, percentages, etc.

Named entity recognition

Apple was founded by Steve Jobs, Steve Wozniak, and Ronald Wayne in April 1976 to develop and sell Wozniak's Apple I personal computer, though Wayne sold his share back to Jobs and Wozniak within 12 days. It was incorporated as Apple Computer, Inc., in January 1977, and sales of its computers, including the Apple II, grew quickly. Apple Inc. headquartered in Cupertino, California.

Named entity recognition

Apple was founded by Steve Jobs, Steve Wozniak, and Ronald Wayne in April 1976 to develop and sell Wozniak's Apple I personal computer, though Wayne sold his share back to Jobs and Wozniak within 12 days. It was incorporated as Apple Computer, Inc., in January 1977, and sales of its computers, including the Apple II, grew quickly. Apple Inc. headquartered in Cupertino, California.

Organization

Person

Location

Time

Named entity recognition

- Common named entities
 - Person
 - E.g., Steve Jobs, Steve Wozniak
 - Organization
 - E.g., Apple, Google, Technische Universität Berlin
 - Time
 - E.g., April 1976, 2006, 16:34, 2am
 - Location
 - E.g., Berlin, California

Named entity recognition

- What are the use cases?
- A lot of relations are associations between named entities

Apple was founded by Steve Jobs, Steve Wozniak, and Ronald Wayne in April 1976

Company	Founders	Founded in
Apple	Steve Jobs Steve Wozniak Ronald Wayne	April 1976
...

Named entity recognition

- What are the use cases?
- A lot of relations are associations between named entities

Apple was founded by Steve Jobs, Steve Wozniak, and Ronald Wayne in April 1976

- For question answering, answers are often named entities
- Sentiment can be attributed to companies or products

I like Google but I hate google chrome

Named entity recognition

- Why is NER difficult?
 - Entity ambiguity
 - Apple produces seeds vs. Apple produces iPhones
- Nested entities
 - University of George Washington

Named entity recognition

- NER is not the only sub-task of information extraction
- Fact Extraction
 - Performs various syntactic transformations on sentences to extract factual information
- Relation Extraction
 - Extract the triplet: predicate, subject, object which will be present in sentences

Apple was founded by Steve Jobs, Steve Wozniak, and Ronald Wayne in April 1976

founded at

Information extraction

- What is information extraction
- Named entity recognition
- **Named entity recognition approaches**
- NER evaluation metrics

Named entity recognition approaches

- NER task
 - Predict entities in a text

Apple was founded by Steve Jobs, Steve Wozniak, and Ronald Wayne in April 1976

Apple	ORG
was	O
founded	O
by	O
Steve	PER
Jobs	PER
Steve	PER
Wozniak	PER
,	O
And	O
Ronald	PER
Wayne	PER
In	O
April	TIME
1976	TIME

Named entity recognition approaches

- Common data standard
 - IOB tagging (Inside–outside–beginning)

Apple was founded by Steve Jobs, Steve Wozniak, and Ronald Wayne in April 1976

Apple	B-ORG
was	O
founded	O
by	O
Steve	B-PER
Jobs	I-PER
Steve	B-PER
Wozniak	I-PER
,	O
And	O
Ronald	B-PER
Wayne	I-PER
In	O
April	B-TIME
1976	I-TIME

Named entity recognition approaches

- Common approaches
 - Rule based NER
 - Sequence models
 - Classical sequence labeling
 - Deep neural sequence models

Rule based NER

- A set of rules is manually crafted by experts to recognize a particular named entity type
- The rules are based on syntactic, linguistic and domain knowledge
 - e.g., a person name often begins with a capital letter

Rule based NER

- Sample rules for detecting person name in text
 - Often consists of a sequence of words each of which begins with a capital letter followed by all lowercase letters (John Ryder)
 - May contain a prefix title such as Mr., Dr. or Prof. (Dr. Enrico Fermi)
 - May contain a suffix such as Jr. or III (as in George Bush Sr.)
 - May contain a designation indicator prefix such as President, Justice, Sen., Colonel or CEO (President Clinton)
 - Does not include special characters such as \$, & or % (Johnson & Johnson)
 - Does not include prepositions (Castle of Windsor)

Rule based NER

- Rule based methods are more applicable in closed domain settings
- Legal text
- Patents

Inventor: Jane Doe

Specification

Title. [Realtime Cloudbased Mobile Web App and Social Rootkit].

Cross References to Related Applications. This application claims the benefit of Applicants' prior provisional application, number [00/000,000], filed on [January 1, 2012].

Field of Invention. The technology relates to the general field of [social media software], and has certain specific application to [haptic bootstrap software-as-a-service].

Background

Summary

The disclosed [THING] does [RESULTS]. It may be used by [EXAMPLES].

Brief Description of the Drawings

Various embodiments of the invention are disclosed in the following detailed description and accompanying drawings.

Fig. 1 illustrates a [three-quarter view of the crankshaft wingnut assembly].

Fig. 2 illustrates an [exploded view of the clockwork linchpin]

Fig. 3 illustrates ...

Sequence models

- Sequence labeling is a type of pattern recognition task that involves the algorithmic assignment of a categorical label to each member of a sequence of observed values
 - Named entity recognition
 - Part of speech tagging
 - Keyphrase extraction

Classical sequence labeling

- The naive approach to this problem is to classify each word independently
 - The main problem with this approach is it assumes that named entity labels are independent which is not the case
 - University of George Washington
- NER is a task that the grammar characterizes interpretable sequences of tags and imposes several hard constraints
 - I-ORG cannot follow B-PER

Classical sequence labeling

- Instead of modeling tagging decisions independently, one should model them jointly
- Conditional Random Field (CRF) is one of the most popular models for sequence tagging
- In CRF, input data and output are sequences and we have to take the previous context into account when predicting on a data point
- we use feature functions that have multiple input values

$$f(s, i, l_{i-1}, l_i)$$

Classical sequence labeling

- Common features:
 - word-count
 - is_capitalized
 - is_stopword
- A sample feature function:
 - The i -th word in the sentence is capitalized return 1 else 0

Apple was founded by Steve Jobs, Steve Wozniak, and Ronald Wayne in April 1976

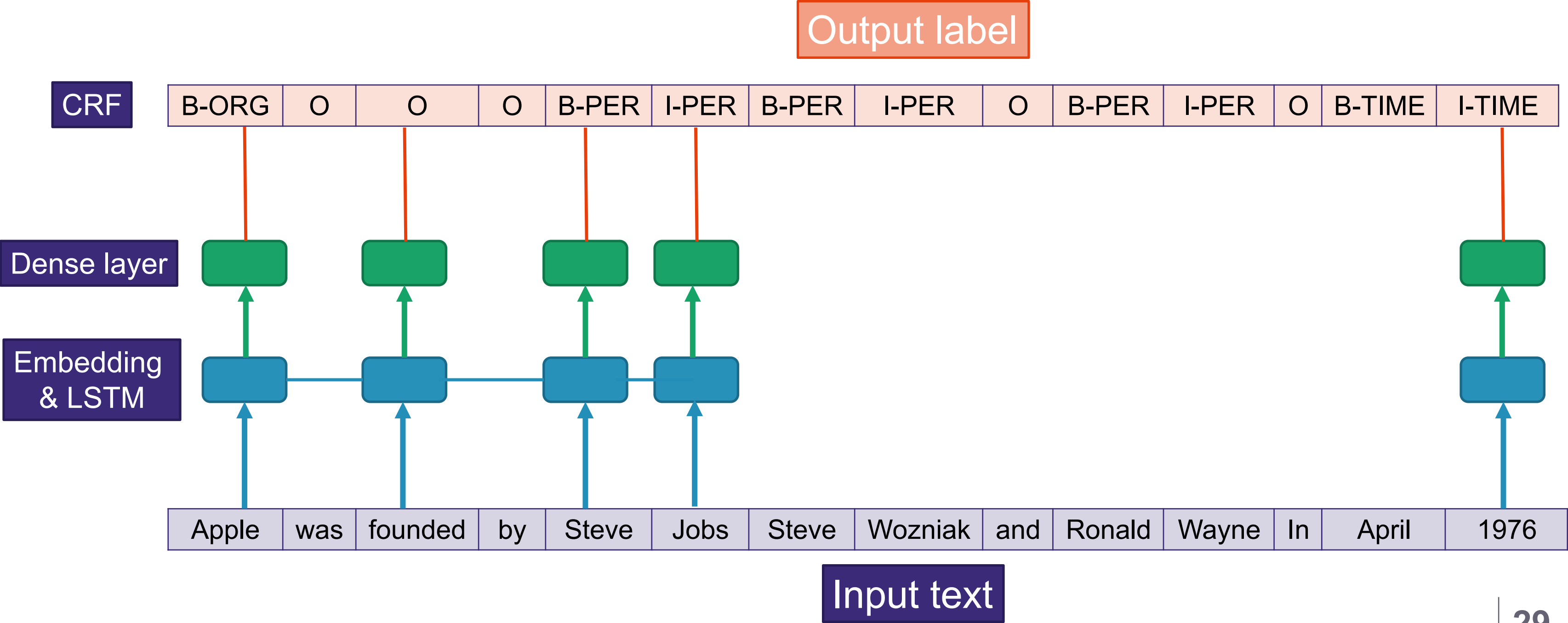
$$f(s, i, l_{i-1}, l_i)$$

$$f(s, 4, 0, \text{PER}) = 0$$

Deep neural sequence models

- Using deep neural networks for tagging a sequence of words
 - Deal with variable-length sequences
 - Maintain sequence order
 - Keep track of long-term dependencies rather than cutting input data too short
 - Share parameters across the sequence (so not re-learn things across the sequence)
- A LSTM can be taken as a Sequence labeler

Deep neural sequence models



Deep neural sequence models

- Sequence tagging using Bi-LSTM (or LSTM) has been explored before where a combination of forward and backward embeddings of each token is passed to a linear classifier
- Produces a probability distribution over all the possible entity-tags for each token
- The CRF layer could add some constraints to the final predicted labels to ensure they are valid
- These constraints can be learned by the CRF layer automatically from the training dataset during the training process

Information extraction

- What is information extraction
- Named entity recognition
- Named entity recognition approaches
- **NER evaluation metrics**

NER evaluation metrics

- Token level vs entity level evaluation
 - Token level metrics are useful to tune a NER system
 - For downstream tasks, it is more useful to evaluate the system with metrics at a full named-entity level

The diagram illustrates the difference between token-level and entity-level Named Entity Recognition (NER) evaluation using the sentence: "Apple was founded by Steve Jobs, Steve Wozniak, and Ronald Wayne in April 1976".

Top Row (Token-level): Each word is highlighted in a separate colored box: "Apple" (red), "was" (dark blue), "founded" (dark blue), "by" (dark blue), "Steve" (magenta), "Jobs," (magenta), "Steve" (magenta), "Wozniak," (magenta), "and" (dark blue), "Ronald" (magenta), "Wayne" (magenta), "in" (dark blue), "April" (green), and "1976" (green). A red box highlights the tokens "Steve Jobs," and another red box highlights the tokens "April 1976".

Bottom Row (Entity-level): The words are grouped into entities: "Apple" (red), "was" (dark blue), "founded" (dark blue), "by" (dark blue), "Steve Jobs," (magenta), "Steve Wozniak," (magenta), "and" (dark blue), "Ronald Wayne" (magenta), "in" (dark blue), "April" (green), and "1976" (green). A red box highlights the entire entity "Steve Jobs," and another red box highlights the entire entity "April 1976".

NER evaluation metrics

- Different potential scenarios regarding the output of a NER system

1. Match (true positive)

Token	Actual label	Predicted label
Apple	B-ORG	B-ORG
was	O	O
founded	O	O
by	O	O
Steve	B-PER	B-PER
Jobs	I-PER	I-PER

The task of sentiment analysis

2. System hypothesized an entity (false positive)

Token	Actual label	Predicted label
was	O	O
founded	O	B-ORG
by	O	O

NER evaluation metrics

3. System misses an entity (false negative)

Token	Actual label	Predicted label
Apple	B-ORG	O
was	O	O
founded	O	O
by	O	O
Steve	B-PER	O
Jobs	I-PER	O

NER evaluation metrics

- Overall performance and performance per entity type

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times (precision \times Recall)}{(precision + Recall)}$$

NER evaluation metrics

4. System assigns the wrong entity type

Token	Actual label	Predicted label
was	O	O
founded	O	O
by	O	O
Steve	B-PER	B-ORG
Jobs	I-PER	I-ORG

NER evaluation metrics

5. Getting a wrong boundaries

Token	Actual label	Predicted label
was	O	O
founded	O	O
by	O	B-PER
Steve	B-PER	I-PER
Jobs	I-PER	I-PER

NER evaluation metrics

6. System assigns the wrong entity type with a wrong boundaries

Token	Actual label	Predicted label
was	O	O
founded	O	O
by	O	B-ORG
Steve	B-PER	I-ORG
Jobs	I-PER	I-ORG

NER evaluation metrics

- Different evaluation schema
- CoNLL: Computational Natural Language Learning
 - Measures the performance of the systems in terms of precision, recall and f1-score
 - A named entity is correct only if it is an exact match of the corresponding entity in the data file

NER evaluation metrics

- Message Understanding Conference (MUC)
 - Correct (COR) : both are the same;
 - Incorrect (INC) : the output of a system and the golden annotation don't match;
 - Partial (PAR) : system and the golden annotation are somewhat “similar” but not the same;
- Missing (MIS) : a golden annotation is not captured by a system;
- Spurious (SPU) : system produces a response which doesn't exist in the golden annotation;

$$Error = \frac{INC + \frac{PAR}{2} + MIS + SPU}{COR + INC + PAR + MIS + SPU}$$

Summary

- Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured text

who invented mouse and when

All

News

Images

Shopping

Videos

More


Settings

Tools

About 11.400.000 results (0,66 seconds)

Douglas Engelbart

The computer mouse was invented and developed by **Douglas Engelbart**, with the assistance of Bill English, during the 1960s and was patented on November 17, 1970. 29 Dec 2017



https://www.computerhope.com > ... > Mouse Help

When and who invented the first computer mouse?

roots book

All

Images

Shopping

Videos

Books

More

Settings

Tools

About 764.000.000 results (0,88 seconds)

https://en.wikipedia.org > wiki > Roots:_The_Saga_of_... >

Roots: The Saga of an American Family - Wikipedia

Roots: The Saga of an American Family is a 1976 novel written by Alex Haley. It tells the story of Kunta Kinte, an 18th-century African, captured as an adolescent, sold into slavery in Africa, transported to North America; following his life and the lives of his descendants in the United States down to Haley.

LC Class: E185 .97.H24 A33

Publisher: Doubleday

Author: Alex Haley

Publication date: August 17, 1976

[Plot](#) · [Characters in Roots](#) · [Reception](#) · [Historical accuracy](#)

People also ask

What happened to Alex Haley?

Was Alex Haley related to Kunta Kinte?

Is roots a true story?

How many sons did Chicken George have?

Feedback

https://www.amazon.com > Roots-American-Family-Ale... >

Roots: The Saga of an American Family: Haley, Alex ...

Based off of the bestselling author's family history, this novel tells the story of Kunta Kinte, who is sold into slavery in the United States where he and his ...

https://www.goodreads.com > book > show >

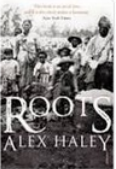
Roots: The Saga of an American Family by Alex Haley

Roots: The Saga of an American Family is a novel written by Alex Haley and first published in 1976. Roots tells the story of Kunta Kinte—a young man taken from the Gambia when he was seventeen and sold as a slave—and seven generations of his descendants in the United States.

★★★★★ Rating: 4.4 · 150,712 votes

Roots: The Saga of an American Family

Novel by Alex Haley



4,4/5 · Goodreads

93% liked this book
Google users

Roots: The Saga of an American Family is a 1976 novel written by Alex Haley. It tells the story of Kunta Kinte, an 18th-century African, captured as an adolescent, sold into slavery in Africa, transported to North America; following his life and the lives of his descendants in the United States down to Haley. [Wikipedia](#)

Originally published: August 17, 1976

Author: Alex Haley

Pages: 704 pp (First edition, hardback)

Awards: [Pulitzer Prize Special Citations and Awards](#)

Adaptations: [Roots \(1977\)](#), [Roots \(2016\)](#), [Roots: The Next Generations \(1979\)](#)

Genres: [Novel](#), [Biography](#), [Historical Fiction](#), [Fictional Autobiography](#)

Book quotes

Characters

Rate and review

Summary

Apple was founded by Steve Jobs, Steve Wozniak, and Ronald Wayne in April 1976 to develop and sell Wozniak's Apple I personal computer, though Wayne sold his share back to Jobs and Wozniak within 12 days. It was incorporated as Apple Computer, Inc., in January 1977, and sales of its computers, including the Apple II, grew quickly. Apple Inc. headquartered in Cupertino, California.

Organization

Person

Location

Time

Summary

- Apple produces seeds vs. Apple produces iPhones
- University of George Washington

Apple.	B-ORG
was	O
founded	O
by	O
Steve	B-PER
Jobs	I-PER
Steve	B-PER
Wozniak	I-PER
,	O
And	O
Ronald	B-PER
Wayne	I-PER
In	O
April	B-TIME
1976	I-TIME

Summary

- Common approaches
 - Rule based NER
 - Sequence models
 - Classical sequence labeling
 - Deep neural sequence models

Apple was founded by Steve Jobs, Steve Wozniak, and Ronald Wayne in April 1976

$$f(s, i, l_{i-1}, l_i)$$

$$f(s, 4, 0, \text{PER}) = 0$$

Inventor. Jane Doe

Specification

Title. [Realtime Cloudbased Mobile Web App and Social Rootkit].

Cross References to Related Applications. This application claims the benefit of Applicants' prior provisional application, number [00/000,000], filed on [January 1, 2012].

Field of Invention. The technology relates to the general field of [social media software], and has certain specific application to [haptic boilerstrap software-as-a-service].

Background

Summary

The disclosed [THING] does [RESULTS]. It may be used by [EXAMPLES].

Brief Description of the Drawings

Various embodiments of the invention are disclosed in the following detailed description and accompanying drawings.

Fig. 1 illustrates a [three-quarter view of the crankshaft wingnut assembly].

Fig. 2 illustrates an [exploded view of the clockwork linchpin]

Fig. 3 illustrates ...

Summary

Output label

CRF

B-ORG	O	O	O	B-PER	I-PER	B-PER	I-PER	O	B-PER	I-PER	O	B-TIME	I-TIME
-------	---	---	---	-------	-------	-------	-------	---	-------	-------	---	--------	--------

Dense layer

Embedding
& LSTM

Apple	was	founded	by	Steve	Jobs	Steve	Wozniak	and	Ronald	Wayne	In	April	1976
-------	-----	---------	----	-------	------	-------	---------	-----	--------	-------	----	-------	------

Input text

Summary

Token	Actual label	Predicted label
was	O	O
founded	O	O
by	O	O
Steve	B-PER	B-ORG
Jobs	I-PER	I-ORG

Token	Actual label	Predicted label
was	O	O
founded	O	B-ORG
by	O	O

Token	Actual label	Predicted label
Apple	B-ORG	O
was	O	O
founded	O	O
by	O	O
Steve	B-PER	O
Jobs	I-PER	O

Token	Actual label	Predicted label
Apple	B-ORG	B-ORG
was	O	O
founded	O	O
by	O	O
Steve	B-PER	B-PER
Jobs	I-PER	I-PER