

Information Extraction

Natural Language Processing Module 2023

Dr Samuel A Mensah
s.mensah@sheffield.ac.uk



The
University
Of
Sheffield.

March 21, 2023

- Information Extraction Background
- System Architecture
- Feature Extraction
- Feature Learning
- Rule- and Machine-learning based methods
- Experimental Methodology
- Information Extraction Tasks:
 - Named Entity Recognition
 - Relation Extraction
- Application: Information Extraction for Question Answering

In this session, we aim to:

- get introduced to information extraction (IE), its concepts and history
- learn about rule based and learning approaches to IE
- application of IE in question answering

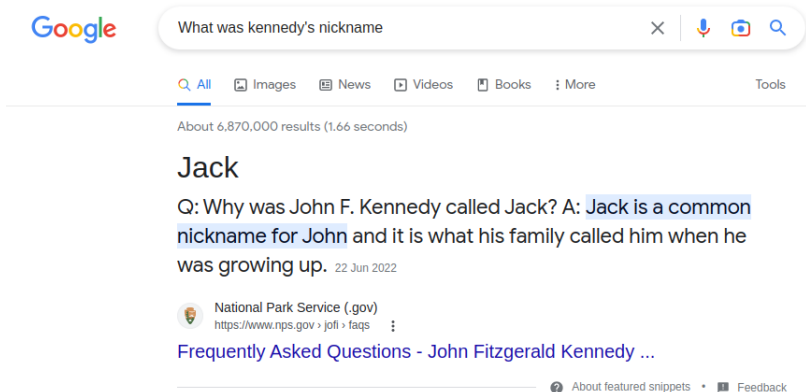
Question:

What was Kennedy's nickname?

Motivation

Question:

What was Kennedy's nickname?



Google


What was Kennedy's nickname

Q All Images News Videos Books More Tools

About 6,870,000 results (1.66 seconds)

Jack

Q: Why was John F. Kennedy called Jack? A: Jack is a common nickname for John and it is what his family called him when he was growing up. 22 Jun 2022

 National Park Service (.gov)
<https://www.nps.gov/jofi/faq>

Frequently Asked Questions - John Fitzgerald Kennedy ...

About featured snippets Feedback

Google Ranking Algorithm: Key factors

- **Meaning of query:** Use language models to try to decipher how the relatively few words match up to content.
- **Relevance of content:** Check whether content is relevant to your query. For example, by examining whether the keywords used in your query can be found within the content.
- **Quality of content:** Prioritise content that seem most helpful. Content demonstrates expertise, authoritativeness and trustworthiness. For example, an application of the Page Rank algorithm.
- **Usability of web pages:** mobile friendly, if content loads quickly, etc.
- **Context and settings:** Country and location to deliver content relevant for your area.¹

¹https://www.google.com/intl/en_uk/search/howsearchworks/how-search-works/ranking-results/

Motivation

ChatGPT:

Model: Default (GPT-3.5)

S

what was kennedy's nickname



John F. Kennedy, the 35th President of the United States, was often referred to by his initials "JFK" or simply as "Jack". He was also known by his family nickname "Johnny" among his close friends and relatives.



“Information Extraction” Defined

What is Information Extraction (IE)?

“Information Extraction” Defined

What is Information Extraction (IE)?

- a practically-motivated engineering discipline (models not necessarily inspired by nature)
- the extraction of structured information from unstructured (= textual) sources.
- its significance is connected to the growing amount of information in text and its potential use in systems (e.g. question answering)

A Short History of Information Extraction

- **1965:** NYU “Linguistic String” project (N. Sager) to facilitate search and retrieval of requested information in scientific and technical literature.
- **1982:** DeJong’s FRUMP (Fast Reading Understanding and Memory Program) system: “sketchy scripts” to highlight key information in text to generate summaries.
- **1987/1989 Message Understanding Conference (MUC) I+II:** IE in naval operation messages sponsored by DARPA. Developing methods and metrics for IE.
- **1991/1998 MUC 3-7:** News reports and other domains
- **2000-2004:** ACE: Automatic Content Extraction
 - defines research objective in terms of the target objects
 - from text spans to abstract entities, e.g., PERSON, LOCATION, COUNTRY
 - extract relations (e.g., capital of), event (e.g., destruction)
 - English, Chinese, Arabic
- **2010s:** First neural approaches to IE
 - CNN , RNN, LSTM, GCN, Transformer [5], BERT [1], GPT [4]

Applications of Information Extraction

There are several application domains for IE systems:

- Bio-medical IE applications (genes and proteins)
- Financial IE applications (e.g., Message Formatting Expert (MFE) system in DBS bank, Singapore to extract relevant information from “letters of credit”.)
- Intelligence & IE applications (terrorists & crimes)
- Reuters News Tracer: enables journalists to spot and validate breaking news in real time on Twitter.
- e-Commerce IE applications (brands & productions, sentiment information extraction for recommender systems)
- **Question Answering** (e.g. Google Search, ChatGPT)

Context

*“John Fitzgerald Kennedy (May 29, 1917 – November 22, 1963), often referred to by his initials JFK and the nickname Jack, was an American politician who served as the 35th president of the United States from 1961 until his assassination in 1963. He was the youngest person to assume the presidency by election and the youngest president at the end of his tenure. **Jack is a common nickname for John and it is what his family called him when he was growing up.**”*

- What was Kennedy's nickname?

Context

“John Fitzgerald Kennedy (May 29, 1917 – November 22, 1963), often referred to by his initials JFK and the nickname Jack, was an American politician who served as the 35th president of the United States from 1961 until his assassination in 1963. He was the youngest person to assume the presidency by election and the youngest president at the end of his tenure. Jack is a common nickname for John and it is what his family called him when he was growing up.”

- What was Kennedy's nickname?
- Who served as the 35th president of the United States?

Context

*"John Fitzgerald Kennedy (May 29, 1917 – November 22, 1963), often referred to by his initials JFK and the nickname Jack, was an American politician who served as the 35th president of the United States from 1961 until his assassination in 1963. **He was the youngest person to assume the presidency by election and the youngest president at the end of his tenure.** Jack is a common nickname for John and it is what his family called him when he was growing up."*

- What was Kennedy's nickname?
- Who served as the 35th president of the United States?
- Who was the youngest to assume the presidency by election?

IE Example: Tasks

By answering these questions, we have implicitly performed a number of IE tasks:

① Named Entity Recognition

- *John Fitzgerald Kennedy, United States*

② Entity Disambiguation

- *Jack is also known as John F Kennedy*

③ Entity Coherence

- *The word “He” makes reference to John F Kennedy*

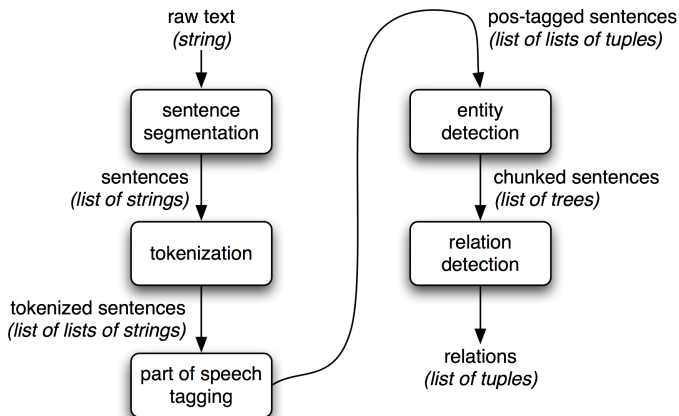
④ Relation Classification/Extraction

- *The phrase “served as the 35th president” shows the relationship between John F Kennedy and United States*
- *Structured knowledge: (John F Kennedy, former_president, United States)*

⑤ Knowledge Base Population

We can automate Question Answering via IE

Information Extraction System Architecture



2

Figure: Pipeline Architecture for an Information Extraction System.

²<https://www.nltk.org/book/ch07.html>

Information Extraction System Architecture

- **Text normalization:** Reducing the text into a single canonical form
- **Tokenization:** Splitting the text into smaller units such as words or characters
- **Stemming:** Reducing inflectional form of words to their base form (e.g., eating, eats, eaten is reduced to eat)
- **Lemmatization:** similar to a stemming process but guided by a lexical knowledge base to obtain accurate word stems.
- **POS tagger:** Assigns part-of-speech tags to words in text
- **Chunk parser:** individual pieces of text and grouping them into meaningful grammatical chunks or syntactic units.
- **Feature Extraction/Learning:** transforming text into numerical features
- **Named entity tagger:** identify and classify entities
- **Relation Tagger:** find relation between entities
- Populate knowledge base with facts

Feature Extraction

Text Classification Example

- **Feature extraction/engineering:** using domain knowledge to extract features from data.
- **Feature:** a piece of evidence intended to help the classifier map the input to the right target class
- **Feature vector:** a vector \vec{F} , the components $F_j = \phi_j(d_j)$, of which are results applying a feature function to the data point d_j .
- **Example:** “Spam vrs Ham” email?
 - number of “!” included in email body
 - length of the email in characters
 - occurrence of the word “cash” in the title or body.
- **Example feature vectors:**
 - (2, 2392, no) → HAM (genuine e-mail)
 - (4, 520, yes) → SPAM
 - (1, 2392, no) → HAM
 - (0, 16337, no) → HAM
 - (0, 61320, yes) → SPAM

Feature Learning

- **Feature learning:** automatically extract features from data - replaces manual feature engineering that is prone to human errors.
- The use of **Supervised** and **Unsupervised learning** techniques for feature learning.
 - Supervised feature learning learns from labelled data and unsupervised learn from unlabelled data.
 - **GloVe vectors** [3] and **word2vec** [2] are vector representations for words that are learned using an unsupervised algorithm.
 - **BERT** [1] is a language model that produces contextual features (i.e., the same word has different representations in different contexts).
 - **Example:** She will park the car so we can walk in the park.
 - **Neural networks** aim to map the input to the output; the hidden-layers of these networks fine-tune input features automatically.

- **Rule-based:**

- Human experts (computational linguists) write general linguistic rules and task-specific extraction rules.
- Example, trigger keywords, regular expressions and patterns.
- Rule-based rules are language dependent, suffer from human ingenuity, time consuming, difficult to adapt to changes.

Regular Expression Example

- Write a regular expression to extract only sheffield or gmail addresses

```
import re
sentence = "My sheffield email address is
           f.surname@sheffield.ac.uk and my
           personal email is f.surname@gmail.com"
reg_exp = "\S+@\ S+"
emails = re.findall(reg_exp, sentence)
```

Machine Learning Methods for IE

Machine-learning based

- **Machine Learning based (supervised):**

- Humans (domain experts) manually annotate text spans indicating entities, relations, facts, etc. in a training corpus;
- features are manually or automatically engineered (or a mixture of the two, e.g. using neural networks and dependency trees);
- these are used to extract information that statistically correlates with classes of entities, relations, etc.

Include:

- Hidden Markov Models (HMMs)
- Conditional Random Fields (CRF)
- Support Vector Machines (SVMs) and Softmax Function
- Artificial Neural Networks (NNs), in particular “deep” neural nets for sequence tagging (RNN, LSTM), CNN, GCN, BERT

Experimental Methodology

- Create a **gold data** - set of reference corpus with **ground truth**
- Split gold data into three parts:
 - **development/training set:** used to study the data, train machine learning processes; can be inspected
 - **development test** ("devtest") set: cannot be inspected, cannot be used for training; repeatedly used to measure improvements of system quality by comparing system output with ground truth.
 - **test set:** cannot be inspected; only used once for final evaluation run at project end. Completely **unseen data** (to the system and developers).
- Gold data split:
could be e.g. 80% train: 10% dev-test : 10% test

Named Entity Tagger

NER: Identify entities (such as person (PER), organisation (ORG), location (LOC), (COUNTRY), (DATE) miscellaneous (MISC)) by assigning tags to chunks in text.

BIOES Encoding:

- **Text:** “John Fitzgerald Kennedy (May 29, 1917 – November 22, 1963), often referred to by his initials JFK and the nickname Jack, was an American politician who served as the 35th president of the United States from 1961 until his assassination in 1963.”
- Annotate word with beginning (B), inside(I), outside(O), end (E), single (S) of class information:
- John (**B-PER**) Fitzgerald (**I-PER**) Kennedy (**E-PER**)
- Jack (**S-PER**)
- United (**B-COUNTRY**) States (**E-COUNTRY**)
- 1963 (**S-DATE**)
- until (**O**)
- Other simpler schemes include BIO

NER Libraries

corenlp.run



version 4.4.0

— Text to annotate —

John Fitzgerald Kennedy (May 29, 1917 – November 22, 1963), often referred to by his initials JFK and the nickname Jack, was an American politician who served as the 35th president of the United States from 1961 until his assassination in 1963.

— Annotations —

named entities X

— Language —

English

Submit

Named Entity Recognition:

	PERSON	DATE	DATE	PERSON	PERSON	NATIONALITY	TITLE	ORDINAL
1	John Fitzgerald Kennedy	(May 29 , 1917 – November 22 , 1963)	, often referred to by his initials	JFK	and the nickname	Jack	, was an American politician who served as the	35th
	TITLE	COUNTRY	DATE	CRIMINAL_CHARGE	DATE			
	president of the United States	from 1961 until his	assassination	in 1963	.			

BiLSTM Approach to NER

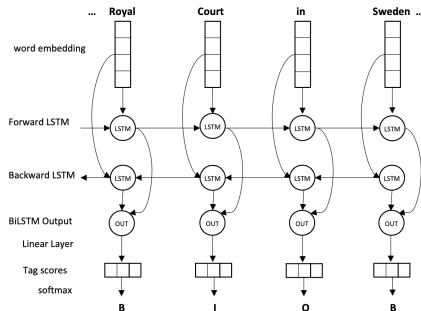


Figure: BiLSTM-Softmax

- BiLSTM-Softmax disregards label dependencies and simply feeds the tag scores into a softmax layer to get the label classification.
- BiLSTM-CRF is more expressive, considers label dependencies, suitable for complex sequence labelling problems such as fine-grained NER (i.e., tags such as {B-ORG, I-ORG} to identify "Royal Court").

John Fitzgerald Kennedy (May 29, 1917 – November 22, 1963), often referred to by his initials JFK and the nickname Jack, was an American politician who served as the 35th president of the **United States** from 1961 until his assassination in 1963. He was the youngest person to assume the presidency by election and the youngest president at the end of his tenure. Jack is a common nickname for John and it is what his family called him when he was growing up.”

John Fitzgerald Kennedy (May 29, 1917 – November 22, 1963), often referred to by his initials JFK and the nickname Jack, was an American politician who served as the 35th president of the **United States** from 1961 until his assassination in 1963. He was the youngest person to assume the presidency by election and the youngest president at the end of his tenure. Jack is a common nickname for John and it is what his family called him when he was growing up.”

- **Relation Extraction:** involves the detection and classification of semantic relationships between pairs of entity mentions.
- **Example 1:** To identify the relation fact (John Fitzgerald Kennedy, former_president, United States)

RE: Representation Learning

Input Features

Text:

***“John Fitzgerald Kennedy** (May 29, 1917 – November 22, 1963), often referred to by his initials JFK and the nickname Jack, was an American politician who served as the 35th president of the **United States** from 1961 until his assassination in 1963.”*

Representation Learning:

- **Word embeddings:** to represent words, i.e., word w_i is mapped to \mathbf{w}_i
- **Relative position embeddings:** to get the relative position of each word to the subject and object entities.
 - the relative position of “president” in the text to “United States” and “John Fitzgerald Kennedy” are -3 and 27, respectively.
 - each relative position p is mapped to a vector \mathbf{d}^p .
 - relative position embeddings for “president” is a concatenated vector $\mathbf{d}_i = [\mathbf{d}^{-3}; \mathbf{d}^{27}]$.
- **Input embeddings:** concatenation of Word and Position Embeddings, i.e., $\mathbf{e}_i = [\mathbf{w}_i; \mathbf{d}_i]$. That is, for the text, we have $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$

1. Sequence-based methods (i.e., BiLSTM, CNN) [7]

- BiLSTM can model **long distance relationships**
- **BiLSTM with attention mechanisms** to capture important features with respect to the relation target [8]. (e.g., “president” is the most important word to determine the relationship between JFK and USA)
- **CNNs** for sequence modelling [6].
- Apply **GNNs** to the dependency trees of text and aggregate with BiLSTM embeddings.
- The use of **BERT** or other language models to learn the text representation for classification.

RE: Model Training

Classification & Training Objective

- We have our final representation \mathbf{v}

$$\mathbf{v} = \text{Model}(\mathbf{E})) \quad (1)$$

- \mathbf{E} denote the input embeddings
- The relation prediction $\hat{y} = \{\hat{y}_1, \dots, \hat{y}_m\}$

$$\hat{y} = \text{softmax}(\mathbf{vW} + \mathbf{b}) \quad (2)$$

- we minimize the cross-entropy loss between the true relation label distribution and the predicted label distribution over all sentences.

$$\mathcal{L} = \sum_s \left(-\frac{1}{m} \sum_{i=1}^m y_i \log \hat{y}_i \right) \quad (3)$$

- \mathbf{E} denote the input embeddings; \mathbf{W}, \mathbf{b} denote learnable weights and bias; m is the number of relations; $y = \{y_1, \dots, y_m\}$ is the one-hot represented ground truth.

Question Answering

An Application

Text:

John Fitzgerald Kennedy (May 29, 1917 – November 22, 1963), often referred to by his initials JFK and the nickname Jack, was an American politician who served as the 35th president of the **United States** from 1961 until his assassination in 1963. He was the youngest person to assume the presidency by election and the youngest president at the end of his tenure. **Jack** is a common nickname for **John** and it is what his family called him when he was growing up.

Relational Facts or Knowledge base:

- (JFK, former_president, USA) , (JFK, nickname, Jack)

Question: What was Kennedy's nickname?

- Disambiguate the entity Kennedy in the query.
- Detect the named entities, JFK, John, Jack in the context.
- Disambiguate the entity John in the context by linking it to JFK
- Detect the relationship between **Jack** and **JFK** in context to extract the fact (JFK, nickname, Jack)
- Answer question using relational fact.

Question Answering

An Application

Question:

What was kennedy's nickname?

The screenshot shows a Google search interface. The search bar contains the text "What was kennedy's nickname". Below the search bar, the Google logo is on the left, and navigation links for "All", "Images", "News", "Videos", "Books", and "More" are in the center. The "All" link is selected. On the right, there are icons for voice search, image search, and a magnifying glass. Below the navigation bar, the search results show "About 6,870,000 results (1.66 seconds)". The first result is titled "Jack" and the snippet reads: "Q: Why was John F. Kennedy called Jack? A: Jack is a common nickname for John and it is what his family called him when he was growing up. 22 Jun 2022". Below the snippet is a link to the National Park Service (.gov) website. At the bottom of the search results, there is a link to "Frequently Asked Questions - John Fitzgerald Kennedy ...".

Google

What was kennedy's nickname

Q All Images News Videos Books More Tools

About 6,870,000 results (1.66 seconds)

Jack

Q: Why was John F. Kennedy called Jack? A: Jack is a common nickname for John and it is what his family called him when he was growing up. 22 Jun 2022

National Park Service (.gov)
<https://www.nps.gov/jofi/faq>

Frequently Asked Questions - John Fitzgerald Kennedy ...

About featured snippets Feedback

Current Research

Multimodal Relation Extraction

Text:

John Fitzgerald Kennedy (May 29, 1917 – November 22, 1963), often referred to by his initials JFK and the nickname Jack, was an American politician who served as the 35th president of the **United States** from 1961 until his assassination in 1963. He was the youngest person to assume the presidency by election and the youngest president at the end of his tenure. **Jack** is a common nickname for **John** and it is what his family called him when he was growing up.

Image:



<https://www.jfklibrary.org/asset-viewer/archives/JFKWHP/>

- Extracting information from images to enhance text-only relation extraction.

In this session, we learned:

- what information extraction is, and a bit of its history.
- the difference between rule-based and machine learning approaches
- how named entities and relations are extracted from text and its application to question answering.

The End

Questions? Comments?

References I



Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.

Bert: Pre-training of deep bidirectional transformers for language understanding.
arXiv preprint arXiv:1810.04805, 2018.



Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean.

Efficient estimation of word representations in vector space.
arXiv preprint arXiv:1301.3781, 2013.



Jeffrey Pennington, Richard Socher, and Christopher D Manning.

Glove: Global vectors for word representation.

In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.



Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al.

Language models are unsupervised multitask learners.
OpenAI blog, 1(8):9, 2019.



Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin.

Attention is all you need.

Advances in neural information processing systems, 30, 2017.



Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao.

Relation classification via convolutional deep neural network.

In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, pages 2335–2344, 2014.

References II



Shu Zhang, Dequan Zheng, Xincheng Hu, and Ming Yang.

Bidirectional long short-term memory networks for relation classification.

In Proceedings of the 29th Pacific Asia conference on language, information and computation, pages 73–78, 2015.



Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu.

Attention-based bidirectional long short-term memory networks for relation classification.

In Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers), pages 207–212, 2016.