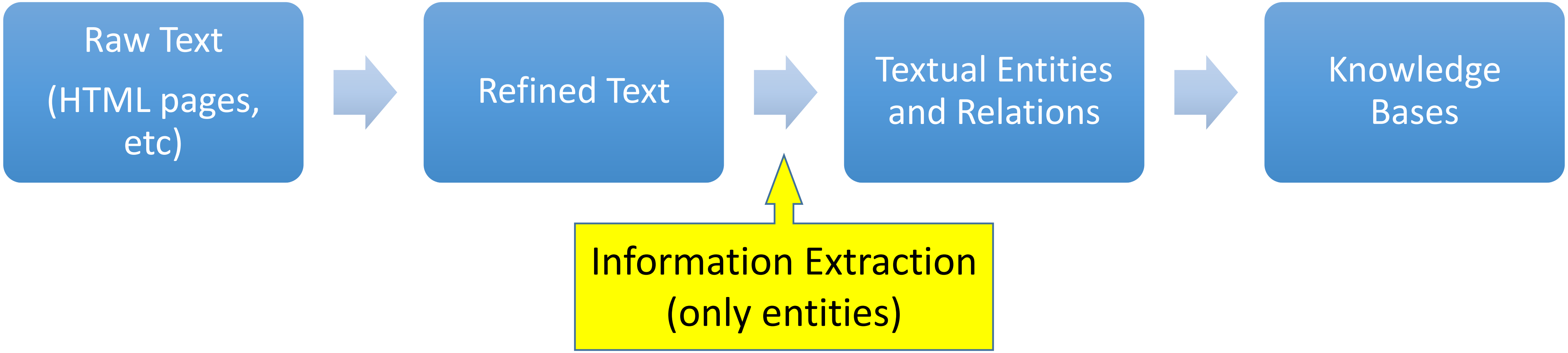
Information Extraction

What is knowledge acquisition?

Knowledge acquisition: process to extract knowledge (to be integrated into knowledge bases) from unstructured text or other data



Raw Text

(HTML pages, etc)

Refined Text

Textual Entities and Relations

Knowledge Bases

Information Extraction (only entities)

Information Extraction

Two types of information extraction

- Named Entity Recognition (NER)
- Relation Extraction (RE)

Sometimes in literature the task of *relation extraction* is referred to as information extraction. We keep the distinction between NER and RE

Rough Accuracy of Information Extraction

Information type	Accuracy
Entities	90-98%
Attributes	80%
Relations	60-70%
Events	50-60%

- These are very rough, actually optimistic, numbers
  - Hold for well-established tasks, but lower for many specific/novel IE tasks

Table from http://web.stanford.edu/class/cs124

Named Entity Recognition (intro)

Slides inspired by http://web.stanford.edu/class/cs124/

Named Entity Recognition (NER)

- **Named entities:** Anything you can refer with a name to with a name
  - Locations, persons, organizations
  - Facilities, vehicles, songs, movies, products
  - Etc.
- **Named entity Recognition:** *find* and *classify* names in text, for example:

“The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.”

Named Entity Recognition (NER)

- **Named entities:** Anything you can refer with a name to with a name
  - Locations, persons, organizations
  - Facilities, vehicles, songs, movies, products
  - Etc.
- **Named entity Recognition: *find*** and *classify* names in text, for example:

“The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.”

Named Entity Recognition (NER)

- **Named entities:** Anything you can refer with a name to with a name
  - Locations, persons, organizations
  - Facilities, vehicles, songs, movies, products
  - Etc.
- **Named entity Recognition:** *find* and ***classify*** names in text, for example:

Person
Date
Location
Organi-
zation

"The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply."

Named Entity Recognition (NER)

- *Detecting entities is useful because*
    - Sentiment can be attributed to companies or products
    - A lot of relations are associations between named entities
    - For question answering, answers are often named entities
    - …

Named Entity Recognition (NER)

- *Like POS, also NER is sequence labeling problem*
- Three standard approaches

1. Hand-written regular expressions
2. Using classifiers
    - Generative: Naïve Bayes
    - Discriminative: Max entropy models
3. Sequence models
    - Hidden Markov Models
    - CMMs (Conditional Markov Models) /MEMMs (Maximum Entropy Models)
    - CRFs (Conditional Random Fields)

Named Entity Recognition (NER)

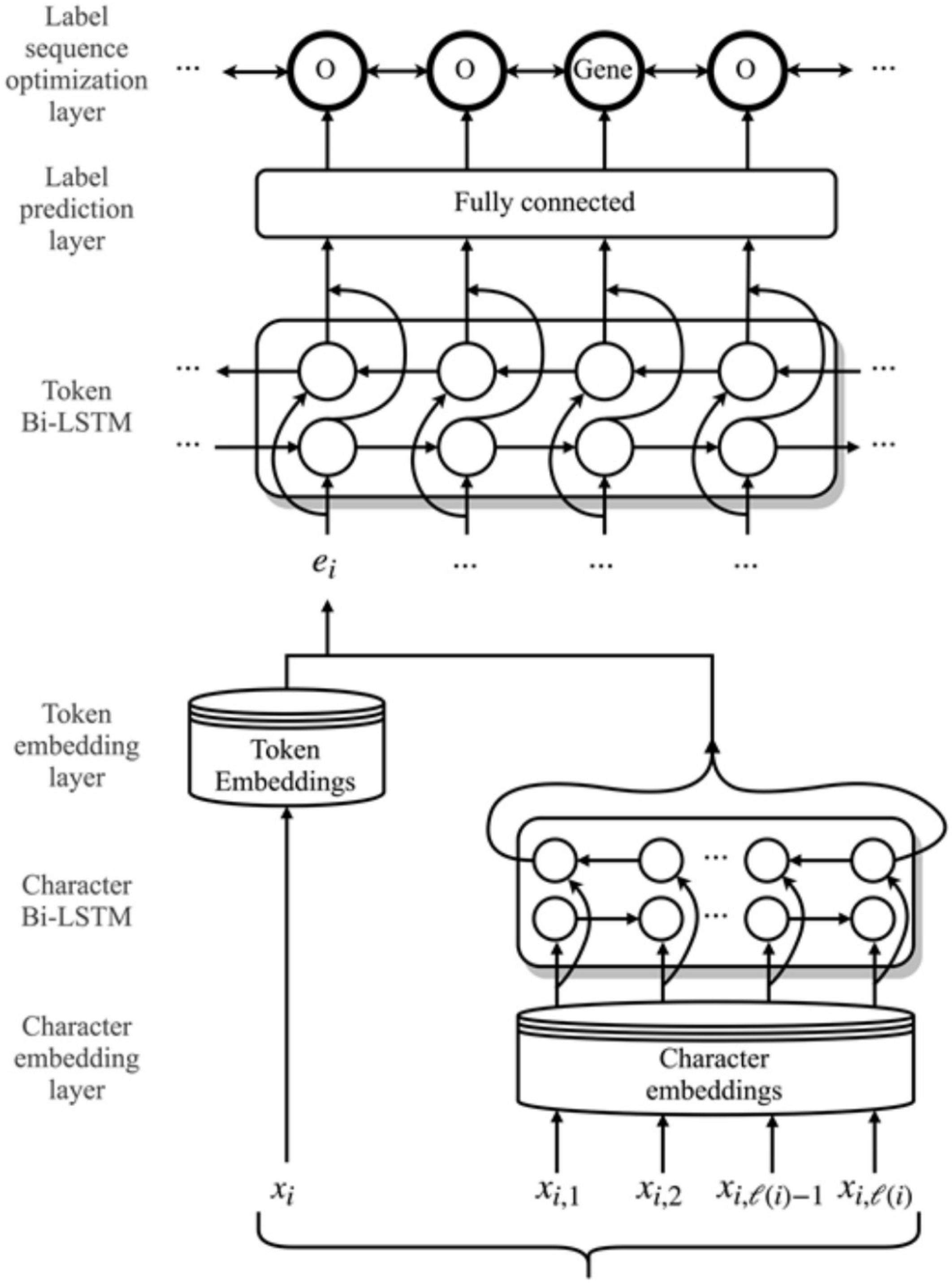
- Typically state-of-the-art Named Entity Recognizers can detect only *gross-grained* types of entities (e.g. Persons, Locations, Organizations)
- Some of the most popular ones
  - Stanford's NER (http://nlp.stanford.edu/software/CRF-NER.shtml) [1]
    - Uses a conditional random fields model
  - Apache OpenNLP (https://opennlp.apache.org/)
    - Uses a maximum entropy model
  - MALLET (http://mallet.cs.umass.edu/)
    - Uses a conditional random fields model
- Recently, large interest in using deep learning models (survey at [2])

[1] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling," in *ACL*, 2005, pp. 363–370.

[2] Yadav, Vikas; Bethand, Steven. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In COLING 2018

NeuroNER [1]

- Uses BiLSTM for entity recognition
- Three layers
  - 1st layer: character-based embeddings concatenated with token-based embeddings
  - 2nd layer: Label prediction layer
  - 3rd layer: Sequence optimization layer



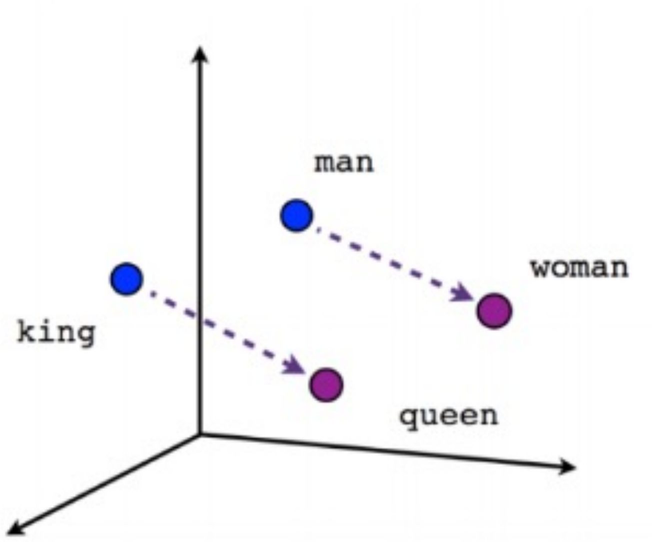
[1] Dernoncourt, Franck; Lee, Ji Young; Szolovits, Peter. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. ACL 2017

Figure from https://academic.oup.com/bioinformatics/article/34/23/4087/5026661

Token (Character) Embeddings

Main idea

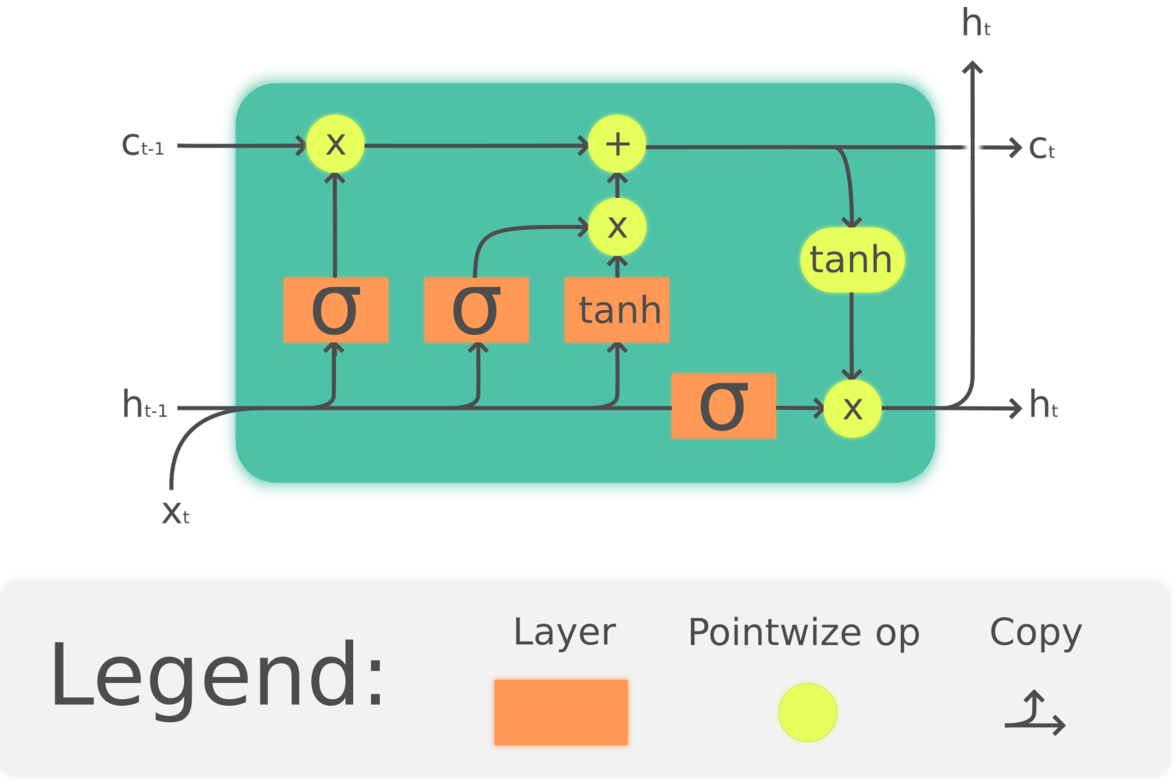
- Map every character, n-gram, token (word) into a vector of real numbers
- Vectors are coordinates into a high-dimensional space
- Learning algorithms compute embeddings so that “similar” words are mapped to “close” vectors



Male-Female

LSTM

- **Long short-term memory (LSTM)** is arguably one of the most famous types of deep learning networks

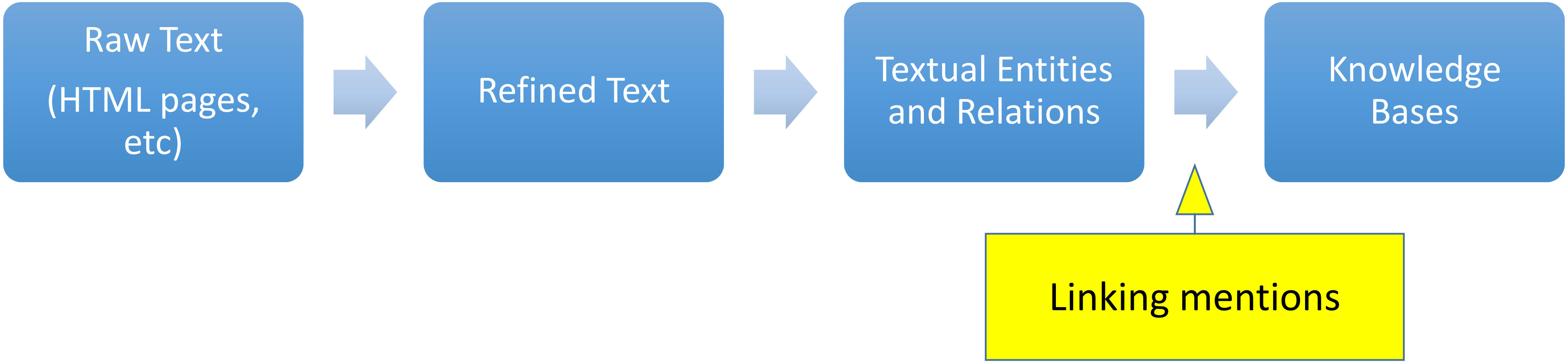


More slides at https://slides.com/laurenstc/lstm
See also http://colah.github.io/posts/2015-08-Understanding-LSTMs/

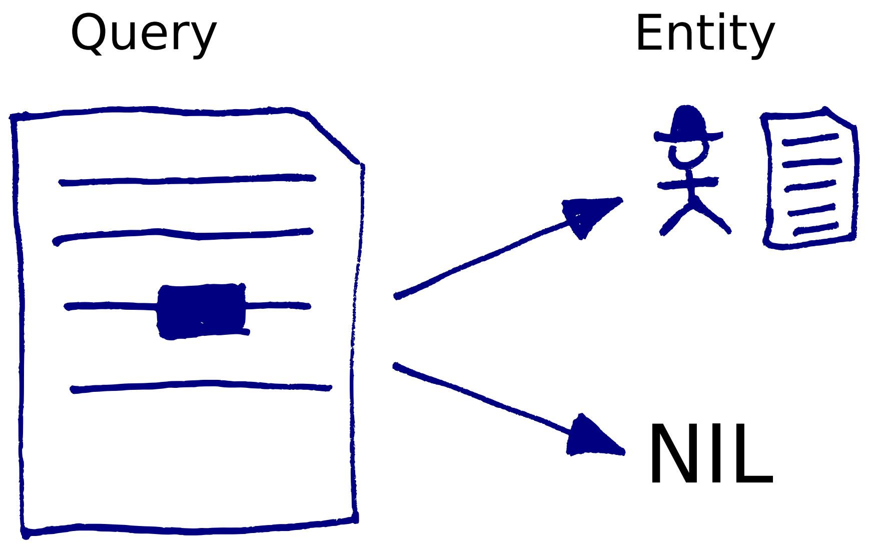
Entity Linking

What is knowledge acquisition?

Knowledge acquisition: process to extract knowledge (to be integrated into knowledge bases) from unstructured text or other data



Problem



Given query mention in a source document,
identify which Wikipedia entity it represents

Problem

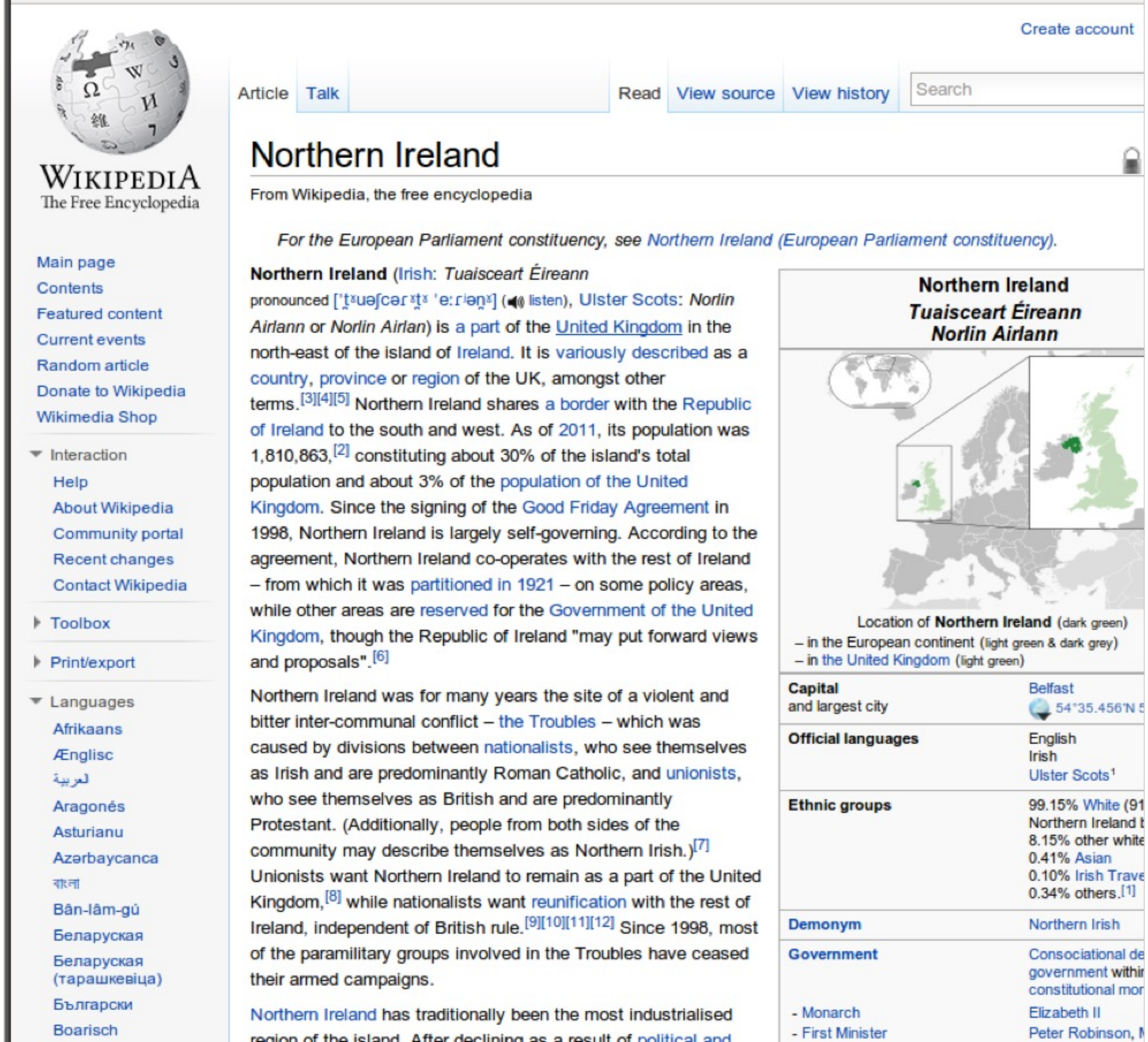
Example Query:

> Northern Ireland has a population of about one and a half million people. At the time of partition in 1921 Protestants / unionists had a two-thirds majority in the region. The first Prime Minister of Northern Ireland, Sir James Craig, described the state as having ‘a Protestant Parliament for a Protestant people.’ The state effectively discriminated against Catholics in housing, jobs, and political representation.
>
> http://cain.ulst.ac.uk/othelem/incorepaper09.htm

Search for:

Northern Ireland

Problem



Northern Ireland

From Wikipedia, the free encyclopedia

For the European Parliament constituency, see Northern Ireland (European Parliament constituency).

Northern Ireland (Irish: *Tuaisceart Éireann* pronounced [ˈt̪ˠuəʃcəɾˠt̪ˠ ˈeːɾʲən̪ˠ] (listen), Ulster Scots: *Norlin Airlann* or *Norlin Airlan*) is a part of the United Kingdom in the north-east of the island of Ireland. It is variously described as a country, province or region of the UK, amongst other terms.[3][4][5] Northern Ireland shares a border with the Republic of Ireland to the south and west. As of 2011, its population was 1,810,863,[2] constituting about 30% of the island's total population and about 3% of the population of the United Kingdom. Since the signing of the Good Friday Agreement in 1998, Northern Ireland is largely self-governing. According to the agreement, Northern Ireland co-operates with the rest of Ireland – from which it was partitioned in 1921 – on some policy areas, while other areas are reserved for the Government of the United Kingdom, though the Republic of Ireland "may put forward views and proposals".[6]

Northern Ireland was for many years the site of a violent and bitter inter-communal conflict – the Troubles – which was caused by divisions between nationalists, who see themselves as Irish and are predominantly Roman Catholic, and unionists, who see themselves as British and are predominantly Protestant. (Additionally, people from both sides of the community may describe themselves as Northern Irish.)[7] Unionists want Northern Ireland to remain as a part of the United Kingdom,[8] while nationalists want reunification with the rest of Ireland, independent of British rule.[9][10][11][12] Since 1998, most of the paramilitary groups involved in the Troubles have ceased their armed campaigns.

Northern Ireland has traditionally been the most industrialised region of the island. After declining as a result of political and

Northern Ireland
Tuaisceart Éireann
Norlin Airlann

Location of **Northern Ireland** (dark green)
– in the European continent (light green & dark grey)
– in the United Kingdom (light green)

Capital and largest city	Belfast 54°35.456′N 5
Official languages	English Irish Ulster Scots¹
Ethnic groups	99.15% White (91 Northern Ireland b 8.15% other white 0.41% Asian 0.10% Irish Trave 0.34% others.[1]
Demonym	Northern Irish
Government	Consociational de government withi constitutional mor
- Monarch	Elizabeth II
- First Minister	Peter Robinson, M

Problem

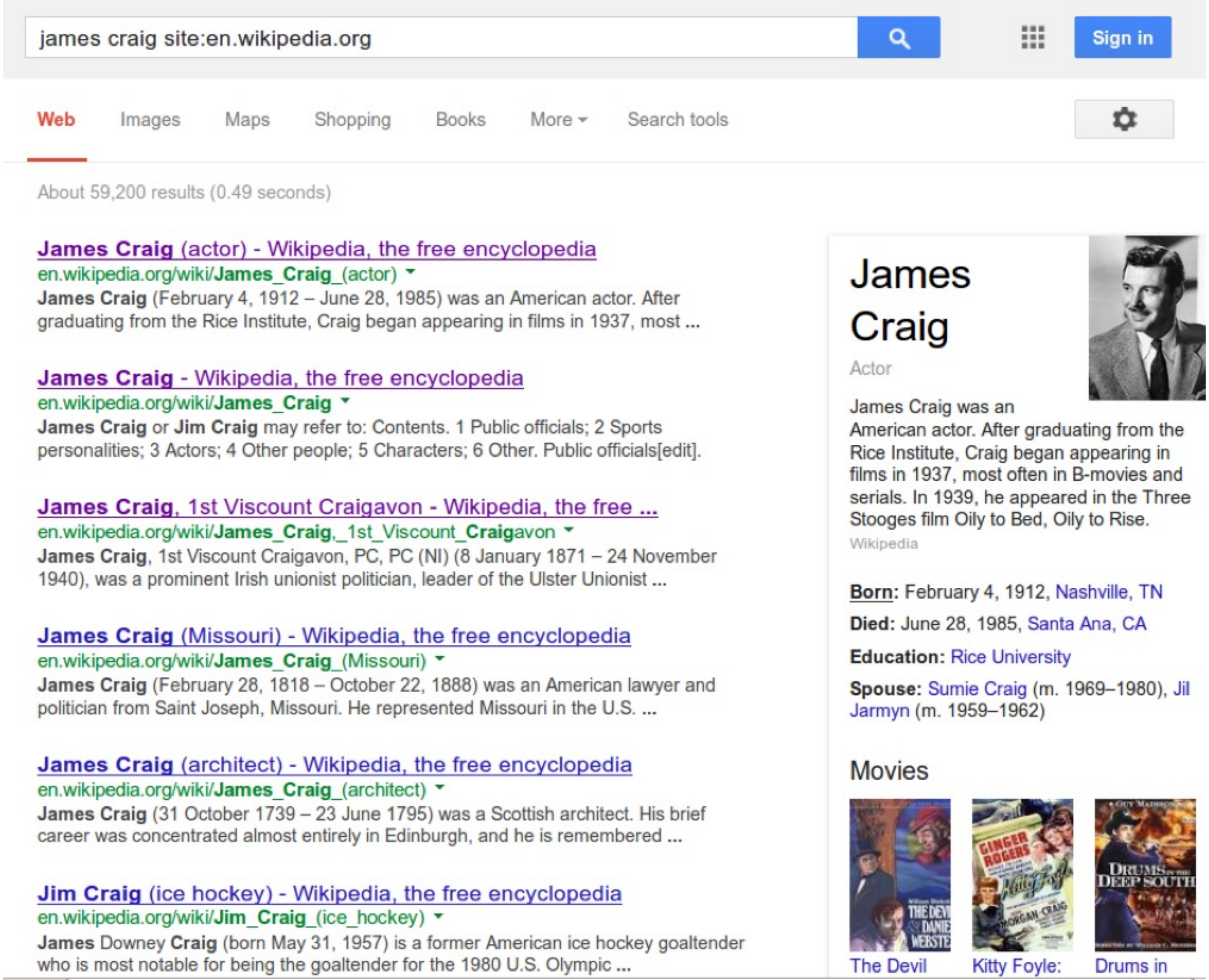
Example Query:

> Northern Ireland has a population of about one and a half million people. At the time of partition in 1921 Protestants / unionists had a two-thirds majority in the region. The first Prime Minister of Northern Ireland, Sir James Craig, described the state as having ‘a Protestant Parliament for a Protestant people.’ The state effectively discriminated against Catholics in housing, jobs, and political representation.
>
> http://cain.ulst.ac.uk/othelem/incorepaper09.htm

Search for:

James Craig

Problem



About 59,200 results (0.49 seconds)

James Craig (actor) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**James_Craig**_(actor)
James Craig (February 4, 1912 – June 28, 1985) was an American actor. After graduating from the Rice Institute, Craig began appearing in films in 1937, most ...

James Craig - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**James_Craig**
James Craig or **Jim Craig** may refer to: Contents. 1 Public officials; 2 Sports personalities; 3 Actors; 4 Other people; 5 Characters; 6 Other. Public officials[edit].

James Craig, 1st Viscount Craigavon - Wikipedia, the free ...
en.wikipedia.org/wiki/**James_Craig**,_1st_Viscount_**Craig**avon
James Craig, 1st Viscount Craigavon, PC, PC (NI) (8 January 1871 – 24 November 1940), was a prominent Irish unionist politician, leader of the Ulster Unionist ...

James Craig (Missouri) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**James_Craig**_(Missouri)
James Craig (February 28, 1818 – October 22, 1888) was an American lawyer and politician from Saint Joseph, Missouri. He represented Missouri in the U.S. ...

James Craig (architect) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**James_Craig**_(architect)
James Craig (31 October 1739 – 23 June 1795) was a Scottish architect. His brief career was concentrated almost entirely in Edinburgh, and he is remembered ...

Jim Craig (ice hockey) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Jim_Craig**_(ice_hockey)
James Downey **Craig** (born May 31, 1957) is a former American ice hockey goaltender who is most notable for being the goaltender for the 1980 U.S. Olympic ...

James Craig

Actor

James Craig was an American actor. After graduating from the Rice Institute, Craig began appearing in films in 1937, most often in B-movies and serials. In 1939, he appeared in the Three Stooges film Oily to Bed, Oily to Rise.
Wikipedia

Born: February 4, 1912, Nashville, TN

Died: June 28, 1985, Santa Ana, CA

Education: Rice University

Spouse: Sumie Craig (m. 1969–1980), Jil Jarmyn (m. 1959–1962)

Movies

Entity Linking: Main Steps

- Task: find *one* mapping from entity **mentions** in the text to **entities** in the Knowledge Base (KB)

- Entity Linking consists of three main operations:
  - **Candidate Entity Generation**
  - **Candidate Entity Ranking**
  - **Unlinkable Mention Prediction**

Candidate Entity Generation

Candidate Entity Generation

- Goal: Link any entity mention to a set of candidates entities in the knowledge base
- Recall is important => If we miss some links, than we will never be able to recover
- **Three main techniques:**
  - Dictionary-based techniques
  - Surface form expansion
  - Based on search-engines

Candidate Entity Generation

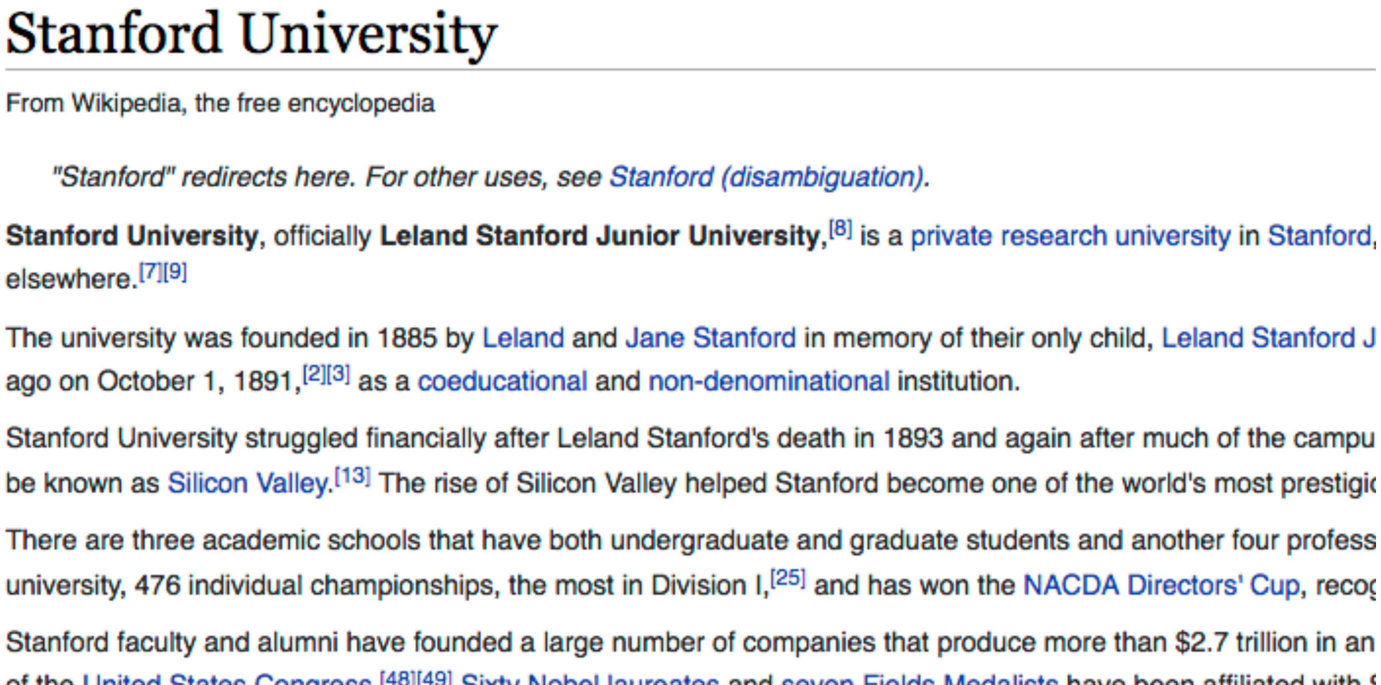
Dictionary-based techniques

- The main idea is to construct an *offline* dictionary *D* between various names and the potential entities in the knowledge base
- **Wikipedia** is the most popular source to construct such vocabulary, so that methods that use it are sometimes referred to as “Wikipedia-based methods”
  - Entity pages
  - Redirect pages
  - Disambiguation pages
  - Bold phrases
  - Hyperlinks

Candidate Entity Generation

Dictionary-based techniques

- Wikipedia entity pages are pages that describe an entity
- The title is typically the name of the entity



Stanford University

From Wikipedia, the free encyclopedia

"Stanford" redirects here. For other uses, see Stanford (disambiguation).

Stanford University, officially **Leland Stanford Junior University**,[8] is a private research university in Stanford, elsewhere.[7][9]

The university was founded in 1885 by Leland and Jane Stanford in memory of their only child, Leland Stanford J ago on October 1, 1891,[2][3] as a coeducational and non-denominational institution.

Stanford University struggled financially after Leland Stanford's death in 1893 and again after much of the campu be known as Silicon Valley.[13] The rise of Silicon Valley helped Stanford become one of the world's most prestigi

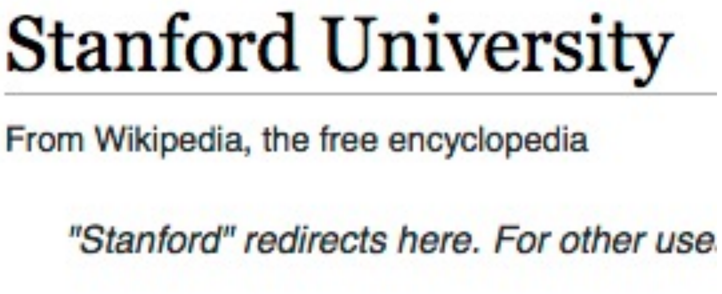
There are three academic schools that have both undergraduate and graduate students and another four profess university, 476 individual championships, the most in Division I,[25] and has won the NACDA Directors' Cup, recog

Stanford faculty and alumni have founded a large number of companies that produce more than $2.7 trillion in an of the United States Congress,[48][49] Sixty Nobel laureates and seven Fields Medalists have been affiliated with S

Candidate Entity Generation

Dictionary-based techniques

- Redirect pages exists for each alternative name for the same identity



Stanford University

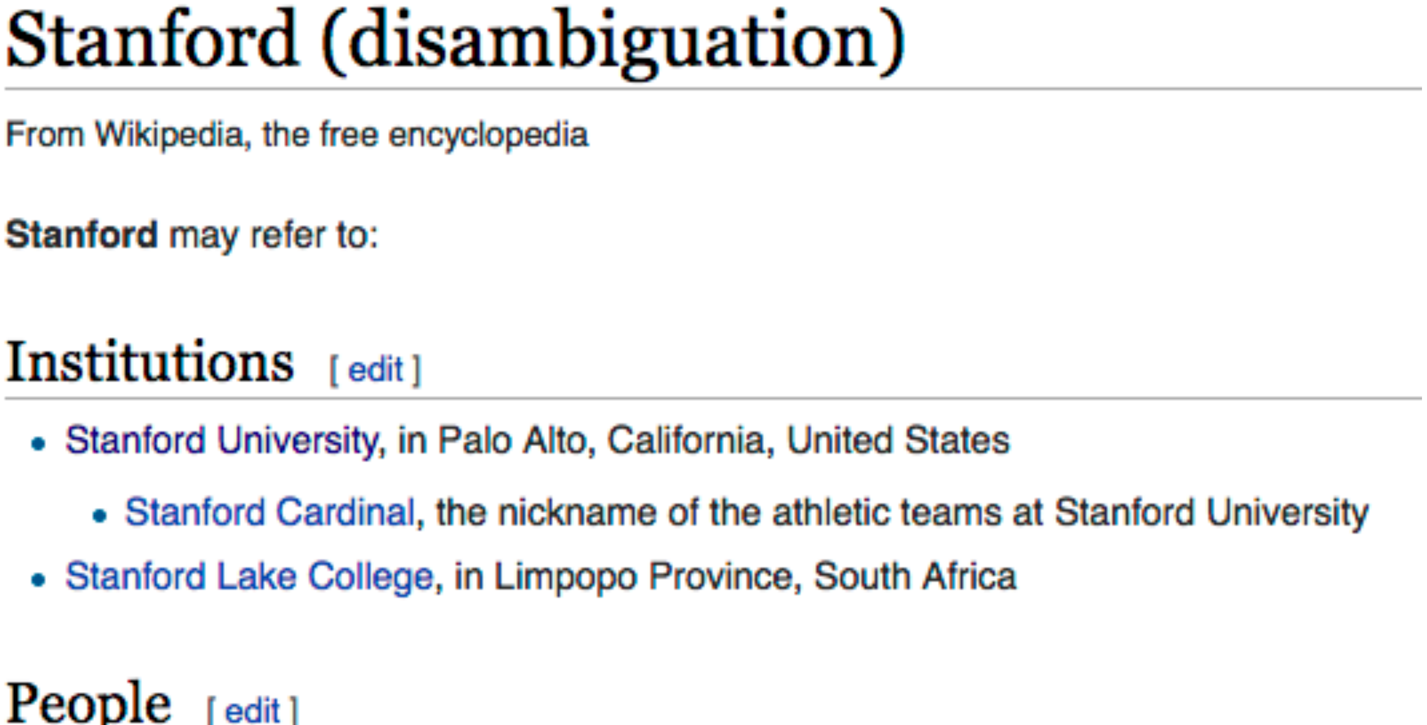
From Wikipedia, the free encyclopedia

"Stanford" redirects here. For other use

Candidate Entity Generation

Dictionary-based techniques

- Disambiguation pages are used when the same name refer to different entities



Stanford (disambiguation)

From Wikipedia, the free encyclopedia

Stanford may refer to:

Institutions [edit]

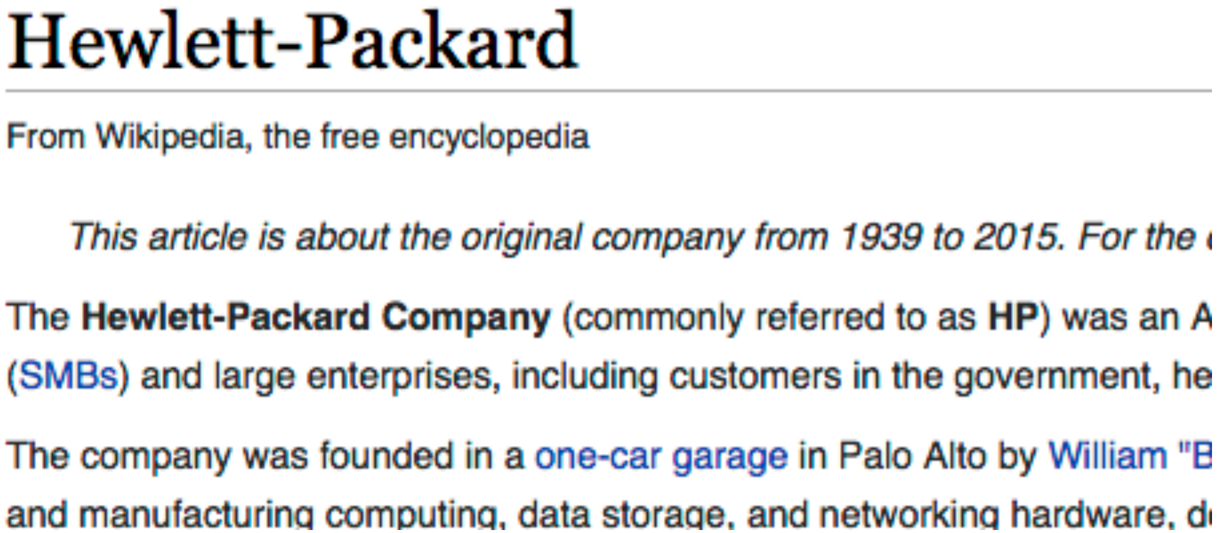
- Stanford University, in Palo Alto, California, United States
  - Stanford Cardinal, the nickname of the athletic teams at Stanford University
- Stanford Lake College, in Limpopo Province, South Africa

People [edit]

Candidate Entity Generation

Dictionary-based techniques

- Bold phrases from the first paragraphs often refers to other names of the same entity



Hewlett-Packard

From Wikipedia, the free encyclopedia

This article is about the original company from 1939 to 2015. For the

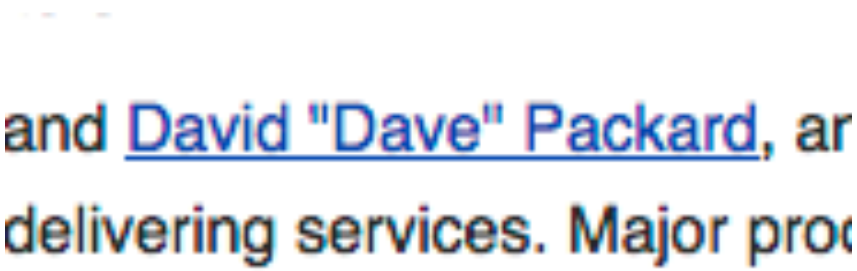
The **Hewlett-Packard Company** (commonly referred to as **HP**) was an A
(SMBs) and large enterprises, including customers in the government, he

The company was founded in a one-car garage in Palo Alto by William "B
and manufacturing computing, data storage, and networking hardware, d

Candidate Entity Generation

Dictionary-based techniques

- Hyperlinks in Wikipedia articles contain anchor text that can be used as synonyms for a particular entity



and David "Dave" Packard, ar
delivering services. Major proc

Candidate Entity Generation

Dictionary-based techniques

- Entities mentions in text can be matched with the dictionary keys either using exact matching or with other string similarity measures
    - Dice score (# common bigrams)
    - Hamming distance (minimum number of substitutions to convert strings)
    - Etc.
- Sometimes entities are misspelled! There are algorithms to correct such cases

Candidate Entity Generation

Surface form expansion

- Some entity mentions are acronyms or part of the entities’ full names
- We can use surface form expansion techniques to identity the real names
  - *Heuristic-based methods*
  - *Supervised learning methods*

Candidate Entity Generation

Surface form expansion – Heuristic-based methods

- We can search the text around the entity mention using pattern matching
    - University of Illinois at Urbana-Champaign (UIUC), HP (Hewlett-Packard)
- We can check for N continuous words (without stop words) that have the same initials than the acronym
- Michael I. Jordan => if “person” according to NER, then also “Jordan” refers to it

Candidate Entity Generation

Surface form expansion – Supervised Learning Methods

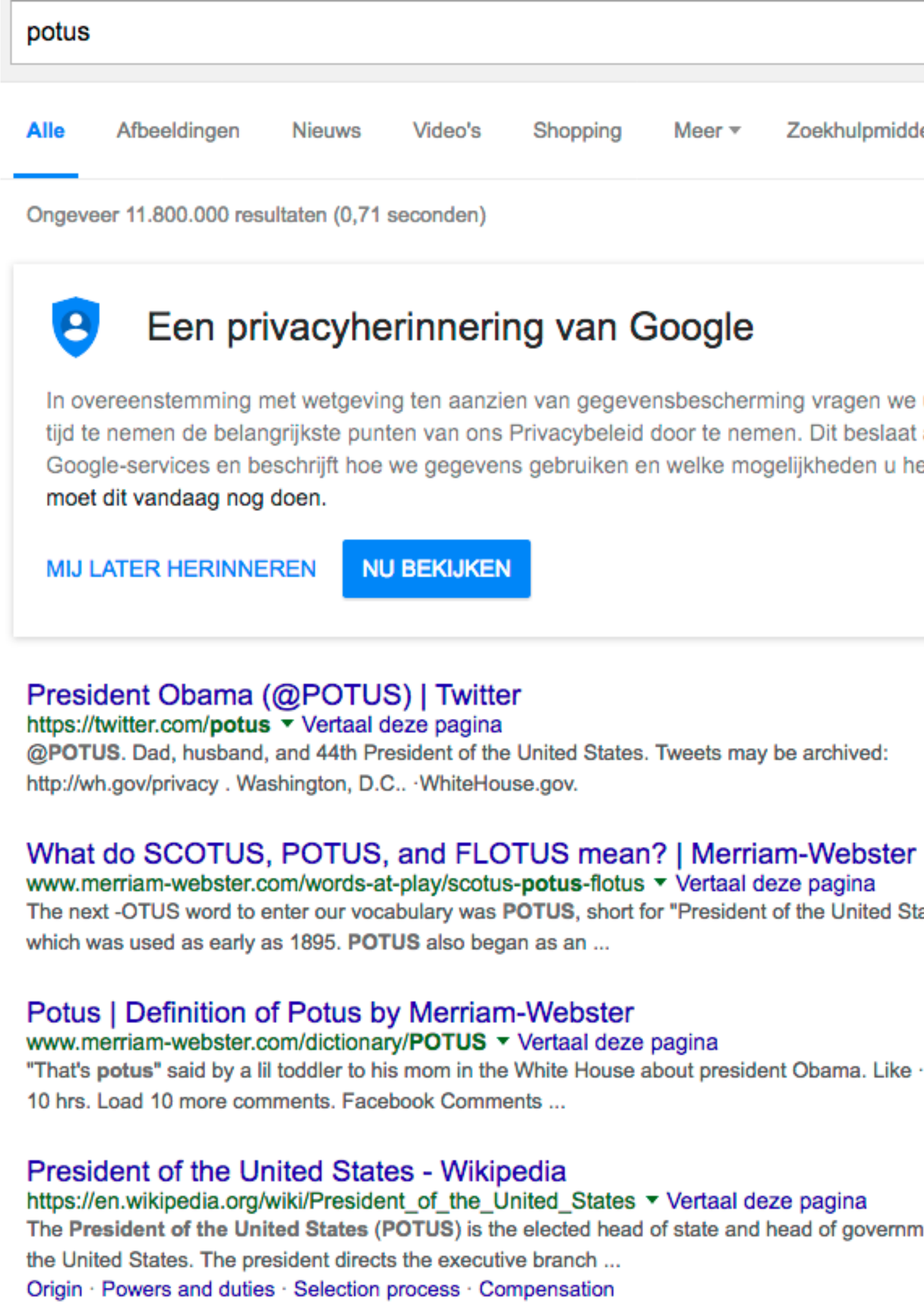
- Some acronyms are quite hard for heuristic based methods
  - E.g. DOD=Department of Defense
- We can use the extractions obtained with heuristics for training a supervised classifier (e.g. SVM). In literature the usage of such classifier improved the accuracy of 15.1% [1]

[1] W. Zhang, Y. C. Sim, J. Su, and C. L. Tan, "Entity Linking with Effective Acronym Expansion, Instance Selection, and Topic Modeling.," in *IJCAI*, 2011, vol. 2011, pp. 1909–1914.

Candidate Entity Generation

Based on search engines

- One simple idea is to query Google with the entity mention and check whether Wikipedia pages show up in the top-k positions
- Disadvantage: this method does not scale!



Ongeveer 11.800.000 resultaten (0,71 seconden)

Een privacyherinnering van Google

In overeenstemming met wetgeving ten aanzien van gegevensbescherming vragen we tijd te nemen de belangrijkste punten van ons Privacybeleid door te nemen. Dit beslaat Google-services en beschrijft hoe we gegevens gebruiken en welke mogelijkheden u he moet dit vandaag nog doen.

President Obama (@POTUS) | Twitter
https://twitter.com/**potus** ▾ Vertaal deze pagina
@**POTUS**. Dad, husband, and 44th President of the United States. Tweets may be archived: http://wh.gov/privacy . Washington, D.C.. ·WhiteHouse.gov.

What do SCOTUS, POTUS, and FLOTUS mean? | Merriam-Webster
www.merriam-webster.com/words-at-play/scotus-**potus**-flotus ▾ Vertaal deze pagina
The next -OTUS word to enter our vocabulary was **POTUS**, short for "President of the United Sta which was used as early as 1895. **POTUS** also began as an ...

Potus | Definition of Potus by Merriam-Webster
www.merriam-webster.com/dictionary/**POTUS** ▾ Vertaal deze pagina
"That's **potus**" said by a lil toddler to his mom in the White House about president Obama. Like · 10 hrs. Load 10 more comments. Facebook Comments ...

President of the United States - Wikipedia
https://en.wikipedia.org/wiki/President_of_the_United_States ▾ Vertaal deze pagina
The **President of the United States** (**POTUS**) is the elected head of state and head of governm the United States. The president directs the executive branch ...
Origin · Powers and duties · Selection process · Compensation