Entity Linking

Laura Dietz dietz@cs.umass.edu
University of Massachusetts

Use cursor keys to flip through slides.
Problem: Entity Linking

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

From Wikipedia, the free encyclopedia

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

Northern Ireland (Irish: Tuaisceart Éireann
pronounced [ˈtuːɪʃərtˠ ˈeːɾʲənˠ] [listen], Ulster Scots: Norlin Airlinn 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, 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 social disruption, the economy has grown and diversified, and the service sector (especially financial and business services) has developed rapidly in recent years.
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
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_Craigavon
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 ...
near miss! :(
Overview

M1: Popularity Method
M2: Machine Learned Similarity
M3: Context with IR
M4: Joint Assignment Model
M5: Joint Retrieval Model

Experimental Results
Online Demos
Challenges

**ambiguity**

/ˌambiˈɡjuəti/  

*noun*

1. uncertainty or inexactness of meaning in language.  
   "we can detect no ambiguity in this section of the Act"  
   *synonyms*: vagueness, obscurity, abstruseness, doubtfulness, uncertainty;  
   More

**synonym**

/ˈsɪnəˌnɪm/  

*noun*

1. a word or phrase that means exactly or nearly the same as another word or phrase in the same language, for example *shut* is a synonym of *close*.  
   *synonyms*: alternate, substitute, alternative, equivalent, [euphemism] More
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
Document Analysis

Name Variants:
   Within-doc Coreference
Neighbor Mentions:
   NER Tagger
   (Alternative Mention Detection)
Sentence:
   Term models

Symbol Notation:

- Q: Query String
- V: Name Variants
- M: Neighbor Mentions
- S: Sentence
Method 1: Popularity of Links

Step 1: Build a dictionary of names for each entity.

Step 2: Inspect all KB entities that have the query mention as a name variant.

Step 3: Choose the entity with the most inlinks through this name.
Names and Links on Wikipedia

Northern Ireland

Northern Ireland (Irish: Tuaisceart Éireann[^4] [twaːʃəɾtˠ ˈɛɾʲən̪ˠ] [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.[^5][^6][^7] Northern Ireland shares a border to the south and west with the Republic of Ireland. As of 2011, its

Unionism in Ireland

In Ireland, unionism also called Ulster unionism in the context of Northern Ireland or - chiefly historically - Ulster specifically, refers to an ideology which favours the continuation of some form of political union between the islands of Ireland and Great Britain or the United Kingdom (UK). Unionists were in the majority in four counties of the Ulster (Antrim, Londonderry, Down and Armagh), and formed a large minority in the remaining counties of Fermanagh and Tyrone. Sir Edward Carson had expressly urged the new Prime Minister of Northern Ireland, Sir James Craig, to ensure absolute equality in the treatment of Catholics, so to guarantee the stability of Unitarianism.

James Craig, 1st Viscount Craigavon

From Wikipedia, the free encyclopedia

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 Party and the first Prime Minister of Northern Ireland. He was created a baronet in 1918 and raised to the Peerage in 1927.

[^4]: [Irish language](https://en.wikipedia.org/wiki/Irish_language)
[^5]: [Northern Ireland](https://en.wikipedia.org/wiki/Northern_Ireland)
[^6]: [Ulster Scots](https://en.wikipedia.org/wiki/Ulster_Scots)
[^7]: [Related terms](https://en.wikipedia.org/wiki/Related_terms)
Pros & Cons: Popularity of Links

Works for very popular entities such as "Northern Ireland"

Fails for entities with confusible names "James Craig", "Springfield", "Jaguar"
Method 1: Popularity of Links

Step 1:
Build a dictionary of names for each entity.

Step 2:
Inspect all KB entities that have the query mention as a name variant.

Step 3:
Choose the entity with the most inlinks through this name.
Method 2: Machine Learn Similarity

Step 1:
Collect different similarity features of query mention and entities

Step 2:
Machine learn the feature weights on training data (e.g. learning to rank)

Step 3:
Apply similarity to query and each entity, select the most similar entity.
Method 2: Similarity Features

- Title: James Craig (actor)
  - Anchor text: James Craig
  - Disambiguation: James Craig
  - Freebase name: James Craig (actor)
- Title: James Craig, 1st Viscount Craigavon
  - Anchor text: Sir James Craig's Craig Administration
  - Disambiguation: James Craig
  - Freebase name: Lord Craigavon
Learn Similarity and NIL

Query

Candidate Entities

Q: Query String
V: Name Variants
M: Neighbor Mentions
S: Sentence

Feature vector for supervised Re-ranking and classification

Re-ranking

1. 2. 3.

NIL classification: Is it similar enough to be a match?

NIL?

Features: Name variants, Document Terms, Links, Popularity ...
Pros & Cons: Machine Learn Similarity

Pro: Combination of different indicators of similarity; option to predict "NILs".

Pro: Can incorporate name variants found in the text (coreference tools).

Con: Requires selection of a pool of candidate entities, which can be large ("John Smith").

Will still fail on "James Craig", because the wrong James has more anchor text matches.
Method 3: Context Disambiguation

Step 1:
Identify surrounding text, entities, etc.

Step 2:
Issue search query containing all of it.
Different Kinds of Context

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 + Name Variants + Neighbors + Sentence
James Craig, 1st Viscount Craigavon - Wikipedia, the free encyclopedia
Member must remember that in the South they boasted of a Catholic State. ... Sir James Craig, Unionist Party, then Prime Minister of Northern Ireland, 24 April ...

Partition of Ireland - Wikipedia, the free encyclopedia
The Prime Minister of Northern Ireland, Sir James Craig, speaking in the ... persecution of Protestants by a Roman Catholic-dominated Irish government.

James Craig - Wikipedia, the free encyclopedia
James Henry Craig (1748–1812), British soldier and colonial administrator; James ... James Craig (County Antrim, 20th century) (1931–1974), Northern Ireland ...

File:James Craig Viscount Craigavon.jpg - Wikipedia, the free encyclopedia
Description, Sir James Craig, 1st Viscount of Craigavon, first Prime Minister of Northern Ireland. Date. Source. This file is lacking source information. Please edit ...

James Craig (loyalist) - Wikipedia, the free encyclopedia
James Pratt "Jim" Craig (c.1941 – 15 October 1988) was a Northern Irish .... Tyrrie suggested that Craig was a suspect on account that his wife was Catholic.

Orange Order - Wikipedia, the free encyclopedia
Catholics, and those whose close relatives are Catholic, are banned from ..... James Craig, the first Prime Minister of Northern Ireland, maintained always that ...

The Troubles - Wikipedia, the free encyclopedia
The conflict was the result of discrimination against the Nationalist/Catholic .... Sir
Method 3: Pros and Cons

Works for "James Craig"!

Problematic when neighbors are ambiguous:

"Lisa witnessed a shooting at Springfield high school".

(Unclear which "Lisa" and which "Springfield")
Method 3: Pros and Cons

Also problematic when neighbors don't provide enough disambiguation power

Example, all other James Craigs of Ireland which are less popular.
James Craig or Jim Craig may refer to:

Public officials

- James Henry Craig (1748–1812), British soldier
- James Craig (County Antrim, 19th century) (1759–1833), British politician
- James Craig (Ontario) (1823–1874), Canadian politician
- James Craig (Missouri) (1818–1888), American politician and soldier
- James Craig (Newcastle) (1834–1902), British politician
- James Craig (University of Dublin) (died 1933), British and Irish politician, physician, and educator
- James Craig, 1st Viscount Craigavon (1871–1940), British politician and Prime Minister of Northern Ireland
- James Craig, 2nd Viscount Craigavon (1906–1974), British politician
- Sir James Craig (diplomat) (born 1924), British ambassador, lecturer, writer
- James Craig (County Antrim, 20th century) (1931–1974), Northern Ireland politician

Contents

1. Public officials
2. Sports personalities
3. Actors
4. Other people
5. Characters
6. Other

Actors

- James Craig (actor) (1912–1985), American actor
- James Philip Craig (actor) (born 1950), British actor

Other people

- James Craig (architect) (1739–1795), Scottish architect
- Sir James Gibson-Craig (1765–1850), 1st Baronet Gibson-Craig
- Sir James Henry Gibson-Craig (1841–1908), 3rd Baronet Gibson-Craig
- James Craig (VC) (1824–1861), British soldier
- James Ireland Craig (1868–1952), British cartographer; creator of Craig retroazimuthal projection
- James Craig (loyalist) (c. 1941–1988), Ulster Defence Association member
- James Humbert Craig (1877–1944), Irish painter

Characters

- Jim Craig (The Man from Snowy River) in films
- Jim Craig (One Life to Live) on television
- James Craig (General Hospital) on television
Method 4: Joint Assignment Models

Step 1: Identify all entity mentions in text
Step 2: For each mention retrieve candidates

Step 3: Select the entity that maximizes:
\[ p(\text{entity}, \text{mention}) \prod (p(\text{entity}), p(\text{entity}, \text{mention})) \]
across all neighbor entities.
Method 4 Example: Candidates

Northern Ireland

James Craig

Catholics

Catholic Church

Roman Catholic Church

American Catholic Church

JC, 1st Viscount Craigavon
Method 4 Example: Correct Selection

James Craig

Northern Ireland

Catholics

Catholic Church

Roman Catholic Church

American Catholic Church

JC, 1st Viscount Craigavon

James Craig (actor)
Method 4 Example: Scoring

Northern Ireland

James Craig

Catholics

American Catholic Church

Method 4 Example: Scoring

Northern Ireland

James Craig

James Craig (actor)

JC, 1st Viscount Craigavon

Catholic Church

Roman Catholic Church

American Catholic Church
Method 4 Example: Wrong Selection

Northern Ireland

James Craig

Catholics

Not compatible

James Craig

James Craig (actor)

Catholic Church

Roman Catholic Church

American Catholic Church

JC, 1st Viscount Craigavon
Method 4: Learn Similarities

As in Method 2, learn feature-based similarity

mention-entity similarity \( p(\text{mention}, \text{entity}) \)

entity-entity similarity \( p(\text{entity}, \text{entity}) \)

entity-entity similarity features:
mutual links, same categories, RDF relations
Method 4: Joint Assignment Models

Step 1: Identify all entity mentions in text
Step 2: For each mention retrieve candidates
Step 3: Select the entity that maximizes:
\[ p(x, z) \prod \left( p(x, y) p(y, z) \right) \]
across all neighbor entities
Method 4: Pros and Cons

Pro: Can mutually resolve uncertainty

Con: Requires a pool of candidates (trade-off runtime versus recall)

Con: Expensive inference problem

May still fail on less popular James Craigs or when context does not resolve ambiguities.
Method 5: Joint Retrieval Model

Step 1: Identify all entity mentions in text
Step 2: For each query mention:
   Issue a search query including query, neighboring mentions, sentence
   Weighting each "ingredient" differently

Intuition: structured matching of text to KB
Northern Ireland

Northern Ireland (Irish: Tuaisceart Éireann[^4] [ˈtuːəʃɛrt 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.[^5][^6][^7] Northern Ireland shares a border to the south and west with the Republic of Ireland. As of 2011, its

Unionism in Ireland

In Ireland, unionism also called Ulster unionism in the context of Northern Ireland or - chiefly historically - Ulster specifically, refers to an ideology which favours the continuation of some form of political union between the islands of Ireland and Great Britain. Unionists were in the majority in four counties of the Ulster (Antrim, Londonderry, Down and Armagh), and formed a large minority in the remaining counties of Fermanagh and Tyrone. Sir Edward Carson had expressly urged the new Prime Minister, Sir James Craig, to ensure absolute equality in the treatment of Catholics, so to guarantee the stability

James Craig, 1st Viscount Craigavon

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 Party and the first Prime Minister of Northern Ireland. He was created a baronet in 1918 and raised to the Peerage in 1927.
James Craig, 1st Viscount Craigavon

Irish Unionist

Unionism in Ireland

Ulster

Northern Ireland

James Craig, 1st Viscount Craigavon

Ulster Unionists

Prime Minister of Northern Ireland

Sir James Craig

1st Viscount Craigavon

Northern Ireland
Connection between 4 and 5

Method 4

\[ p(\text{Q, V}) \Pi (p(\text{O, V}) p(\text{V, M})) \]

Requires iterative optimization

Method 5

Integrate over \text{V}

\[ p(\text{Q, V}) \Pi (p(\text{O, V})) \]

Can be solved inside a search engine
Need a Search Index for the KB

Identify context of query mention

Preprocessing:
build a special KB Index

neighbor-entity similarity features:
neighbor occurs in entity's text
neighbor is title of inlinks/outlinks
Neighbor-Entity Features

- Neighbor occurs in text?
- Neighbor in inlink titles?
- Neighbor in outlink titles?
- Is approx match?
- TF-IDF similarity score

Machine learn the feature weights on training data (e.g. learning to rank)
Query mention-Entity Features

Machine learn the feature weights on training data (e.g. learning to rank)
Method 5: Joint Retrieval Model

Issue the Entity Linking IR

Actually: structured matching

with special KB Index

Select the entity that maximizes:

\[ p(\text{IR}, \text{KB}) \prod p(\text{IR}, \text{Entity}) \]
Method 5: Pros and Cons

Pro: Similar to joint assignment, but cheaper

Pro: Does not require pools (optimize in IR)

Pro: Can be combined with Machine Learning (Method 2) to improve precision.

Con: Fails when context is misleading
Really Difficult Example

Example Query:

**ABC** shot "Lost" in **Australia**

True entity:

American Broadcasting Company

Context "Australia" and mention similarity will point instead to

Australian Broadcasting Corporation

Approach: Identify misleading neighbors (variant of M5)
TAC KBP Entity Linking Task

average recall

cutoff rank $k$

---

<table>
<thead>
<tr>
<th>Year</th>
<th>Graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td><img src="image" alt="2009 Graph" /></td>
</tr>
<tr>
<td>2010</td>
<td><img src="image" alt="2010 Graph" /></td>
</tr>
<tr>
<td>2011</td>
<td><img src="image" alt="2011 Graph" /></td>
</tr>
<tr>
<td>2012</td>
<td><img src="image" alt="2012 Graph" /></td>
</tr>
</tbody>
</table>

- **Q**: Query String
- **V**: Name Variants
- **M**: Neighbor Mentions
- **S**: Sentence

Legend:
- M1 (Popularity)
- M5 + M2 (JR + ML)
- QVM_nrm (Joint Retrieval)
- QVM_nrm LTR
- Q: Query String
- V: Name Variants
- M: Neighbor Mentions
- S: Sentence
References

M1 Popularity / Keyphraseness:
  Wikify!: linking documents to encyclopedic knowledge.

M2 Machine Learn Mention-to-Entity Similarity
  "Using Encyclopedic Knowledge for Named entity Disambiguation."
  "Entity disambiguation for knowledge base population".

M4 Joint Assignment
  "Large-scale named entity disambiguation based on wikipedia data."
Ratinov et al. ACL 2011.
  "Local and global algorithms for disambiguation to wikipedia."
  "Learning relatedness measures for entity linking."

M5 Joint Retrieval Model
  "A neighborhood relevance model for entity linking."

more: http://nlp.cs.rpi.edu/kbp/2014/elreading.html
http://www.mendeley.com/groups/3339761/entity-linking-and-retrieval/
Toolkits & Online Demos

List of toolkits:
http://nlp.cs.rpi.edu/kbp/2014/tools.html

Several Online Demos:
UIUC Wikifier
http://cogcomp.cs.illinois.edu/demo/wikify/

TagMe!
http://tagme.di.unipi.it/

AIDA
https://gate.d5.mpi-inf.mpg.de/webaida/
UIUC Wikifier
TagMe!
AIDA (prior+sim+coherence)
AIDA (prior only)
Another Example: Lisa Fletcher

In 1990, Fletcher began her on-air career while still in college with KEZI-TV in Eugene, Oregon.[3] After graduating, she accepted a job anchoring daily newscasts for KTVZ-TV in Bend, Oregon. In 1993, she returned to Eugene and KEZI where she was the main anchor and investigative reporter. She covered natural disasters, including some of the Northwest's most devastating floods and forest fires. She also anchored coverage of one of the nation's worst high school shootings, the May 1998 tragedy at Thurston High School in Springfield, Oregon, adjacent to Eugene. The coverage she anchored of the Thurston High shootings was honored with an Edward R. Murrow Award for broadcast excellence.[4]
UIUC Wikifier
For instance, you can compute semantic relatedness between topics identified by TAGME, or enable the special parser for Twitter messages. Check the RESTful API page for further details.

Developed by Paolo Ferragina and Ugo Scalzi at A² Lab Dipartimento di Informatica, University of Pisa.

Lisa also anchored coverage of one of the nation's worst high school shootings, the May 1998 tragedy at Thurston High School in Springfield, Oregon.

Lisa Simpson
Lisa Marie Simpson is a fictional main character in the animated television series The Simpsons. She is the middle child of the Simpson family. Voiced by Yeardley Smith, Lisa first appeared on television...

Lisa also anchored coverage of one of the nation's worst high school shootings, the May 1998 tragedy at Thurston High School in Springfield, Oregon.

en.wikipedia.org/wiki/Lisa_Simpson
AIDA
Search Engine (DuckDuckGo)

Lisa Fletcher - Wikipedia, the free encyclopedia
She also anchored coverage of one of the nation's worst high school shootings, the May 1998 tragedy at Thurston High School in Springfield, Oregon, adjacent to Eugene.
[en.wikipedia.org/wiki/Lisa_Fletcher](en.wikipedia.org/wiki/Lisa_Fletcher) More from en.wikipedia.org

Lisa Fletcher named ABC News correspondent - FishbowlNY
She also anchored coverage of one of the country's worst high school shootings that took place in Springfield, Oregon. The coverage was honored with an Edward R. Murrow award for broadcast excellence.
[mediabistro.com/fishbowlny/lisa-fletcher-named-abc-news](mediabistro.com/fishbowlny/lisa-fletcher-named-abc-news) More from mediabistro.com

School massacre - Free net encyclopedia - Main Page - Free ...
In the worst school massacre in United States history, the Bath School disaster, the perpetrator was a member of the ... May 21, 1998: Springfield, Oregon: ... killed and more than 20 other students were injured when 15-year-old Kip Kinkel opened fire in the school cafeteria of Thurston High School.

School massacre
A thorough study of all U.S. school shootings by the U.S. Secret Service warned against the belief that a certain "type" of student would be a perpetrator. ... May 21, 1998: Springfield, Oregon, USA: Kip Kinkel opened fire with a .22
Participate!

TAC KBP Entity Linking Task
http://nlp.cs.rpi.edu/kbp/2014/

SIGIR Entity Recognition and Disambiguation Challenge

INEX 2014 Tweet Contextualization Track
https://inex.mmci.uni-saarland.de/tracks/qa/

Questions?
email: dietz@cs.umass.edu
web: http://ciir.cs.umass.edu/~dietz/