Engaging Content
Engaging People

Constructing a Knowledge Base for Entity Linking on Irish Cultural Heritage Collections

Gary Munnelly

The ADAPT Centre is funded under the SFI Research Centres Programme(Grant 13/RC/2106)and is co-funded under the European Regional Development Fund.

Motivation

The Online Depositions Website

Fully searchable digital edition of the 1641 Depositions at Trinity College Dublin Library, comprising transcripts and images of all 8,000 depositions, examinations and associated materials in which Protestant men and women of all classes told of their experiences following the outbreak of the rebellion by the Catholic Irish in October, 1641... More

Using the 1641 Depositions

Motivation

How do we:

- organise
- facilitate access to
- inform users about the contents of
- enrich/enhance

digital cultural heritage collections?

Problem

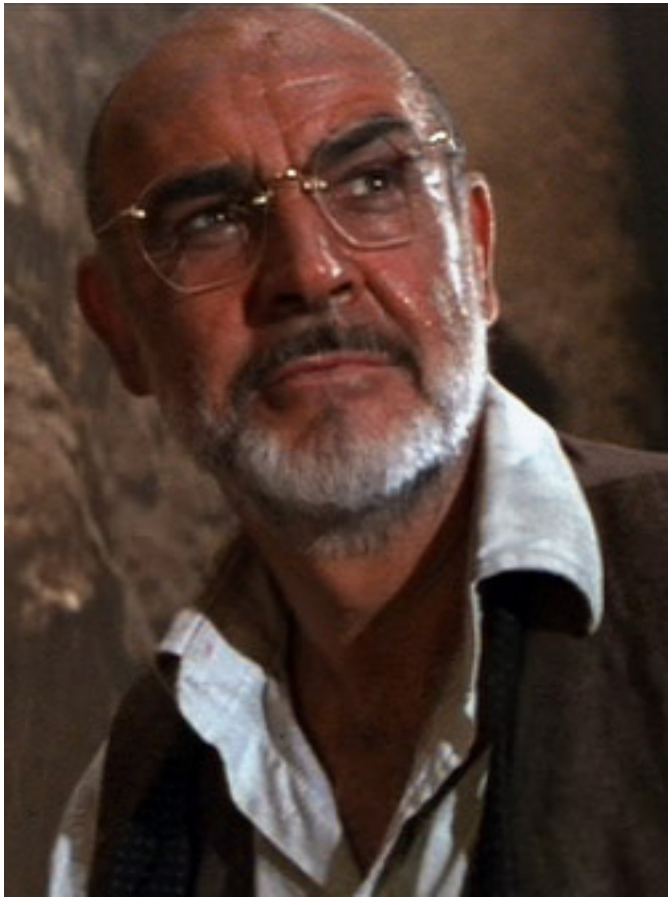
Given a set of ambiguous surface forms, identify the referents.

Entity Linking

I Henry Jones Doctor in Diuinity in obedience to his maiesties Comission requireing an accompt of the losses of his lojall subjects wherein they suffered by the present Rebellion in Ireland...

Dr. Henry Jones?

- Dr. Henry Jones
- Archaeologist
- Possible zombie



Dr. Henry Jones?

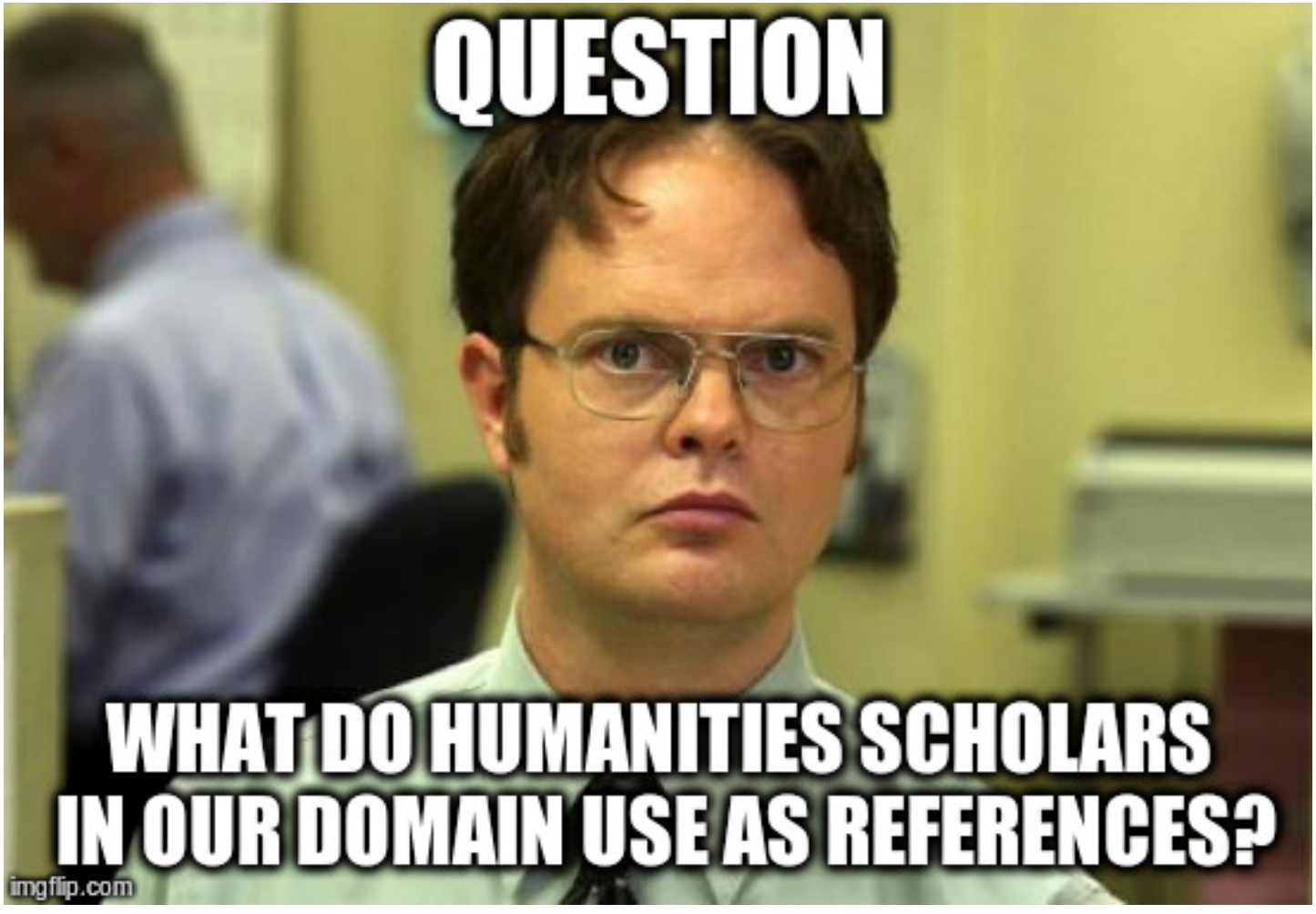
- Dr. Henry Jones
- Bishop of Meath
- Leader of the Commission for the Despoiled Subject

Problems

Finding a good knowledge base can be hard:

- Poor coverage of historical entities in knowledge bases.
- Modern entities can skew EL away from historical ones.
- Lack of trust surrounding Wikipedia based knowledge bases.
- Low tolerance for error.

Question


QUESTION
WHAT DO HUMANITIES SCHOLARS IN OUR DOMAIN USE AS REFERENCES?
imgflip.com

Primary Sources

- Statute Staple
- Petty Maps
- Books of Survey and Distribution
- Lodges Peerage

Primary Sources

- Statute Staple
- Petty Maps
- Books of Survey and Distribution
- Lodges Peerage

Excellent sources of information, but very isolating. Essentially reduces our efforts to little more than record resolution. Guaranteed to be noisy and open to dispute.

Secondary Sources

- Oxford Dictionary of National Biography (ODNB)
- Dictionary of Irish Biography (DIB)

Secondary Sources

- Oxford Dictionary of National Biography (ODNB)
- Dictionary of Irish Biography (DIB)

Not bad! Coverage is not as good as the primary sources, but entities are more concrete.

Still somewhat isolating. Can we connect these resources to a more established knowledge base?

Bonus Benefits!

- More trustworthy information sources
- Limits the Knowledge Base to a specific geographic region

What Information Do We Need?

There are many different types of Entity Linker.

Commonly used features:

- Basic statistical information
  - surface form similarity
  - entity popularity
  - probability that anchor text links to resource
- Relationships with other candidate referents
- Contextual similarity with source text

What Does A Biography Look Like?

Googe, Barnabe

by Judy Barry and Terry Clavin

Googe, Barnabe (*c.*1538–94), poet and translator, and provost-marshal to the presidency court in Connacht, was son of Robert Googe, recorder of Lincoln, and his wife Margaret, daughter of Sir Walter Mantell. He was born at Alvingham, Lincolnshire, and educated at Christ's College, Cambridge, and New College, Oxford. He matriculated at Christ's College in May 1555, before leaving university without graduating in 1556. In 1560 he was a member of Staples Inn, but did not pursue his legal studies, preferring to become a poet and translator. He had published a poem in 1559, and from 1560 started to translate the work of the poet Marcellus Palingenius. Entitled *The Zodiac of life*, it was brought out in a series of twelve books, the first three of which were published that year. The work was finally completed in 1565 and was widely admired. A further work of his, *Eglogs, Epytaphes and Sonettes*, was published in 1561, copies of which are preserved in the Huth, Caell, and Britwell libraries. In 1570 he published the work by which he is best known, a translation into English of Thomas Kirchmeyer's *The popish kingdome, or reigne of Antichrist*, a Latin work which reflected Googe's puritan sympathies.

What Does A Biography Look Like?

Googe, Barnabe

by Judy Barry and Terry Clavin

Googe, Barnabe (*c.*1538–94), poet and translator, and provost-marshal to the presidency court in Connacht, was son of Robert Googe, recorder of Lincoln, and his wife Margaret, daughter of Sir Walter Mantell. He was born at Alvingham, Lincolnshire, and educated at Christ's College, Cambridge, and New College, Oxford. He matriculated at Christ's College in May 1555, before leaving university without graduating in 1556. In 1560 he was a member of Staples Inn, but did not pursue his legal studies, preferring to become a poet and translator. He had published a poem in 1559, and from 1560 started to translate the work of the poet Marcellus Palingenius. Entitled *The Zodiac of life*, it was brought out in a series of twelve books, the first three of which were published that year. The work was finally completed in 1565 and was widely admired. A further work of his, *Eglogs, Epytaphes and Sonettes*, was published in 1561, copies of which are preserved in the Huth, Caell, and Britwell libraries. In 1570 he published the work by which he is best known, a translation into English of Thomas Kirchmeyer's *The popish kingdome, or reigne of Antichrist*, a Latin work which reflected Googe's puritan sympathies.

We use a combination of DBpedia ontology, FOAF and CIDOC-CRM to create ontology.

Identifying Surface Forms

Butler (le Botillier, Pincerna), Theobald

- Theobald Butler
- Theobald le Botiller
- Theobald Pincerna

Identifying Surface Forms

Butler (le Botillier, Pincerna), Theobald

- Theobald Butler
- Theobald le Botiller
- Theobald Pincerna

Bradley, Daniel Joseph ("Dan")

- Daniel Joseph Bradley
- Dan Bradley
- Dan

Identifying Surface Forms

Butler (le Botillier, Pincerna), Theobald

- Theobald Butler
- Theobald le Botiller
- Theobald Pincerna

Bradley, Daniel Joseph ("Dan")

- Daniel Joseph Bradley
- Dan Bradley
- Dan

Anchor text?

Lifespan

- Extracting years of birth/death can be useful for bounding Entity Linking.
- Not all biography dates are concrete e.g. (1609/13 – 1642)
- Use CIDOC-CRM to model probabilistic dates

Identifying Relationships

For now we just use hyperlinks between biographies.

There is a lot of information in the first paragraph of a biography. It would be good to extract that.

Mapping Biographies To DBpedia

Mapping the resulting Knowledge Base to DBpedia is a good idea:

- Can avail of information in DBpedia when it is available
- Enables us to integrate with collections that use DBpedia as vocabulary
- Approaches goal of EL systems which can link over multiple Knowledge Bases

Mapping Biographies To DBpedia

1. Index DBpedia article content and labels in search index
2. Execute biography title as query against index
   - Take top 10 results
3. Compute similarity between biography content and each candidate's content
4. Select candidate with most similar surface form & content.

Mapping Biographies To DBpedia

Final similarity was computed as:

$$sim(a, b) = 0.1 \times sim(a_{sf}, b_{sf}) + 0.9 \times sim(a_{content}, b_{content})$$

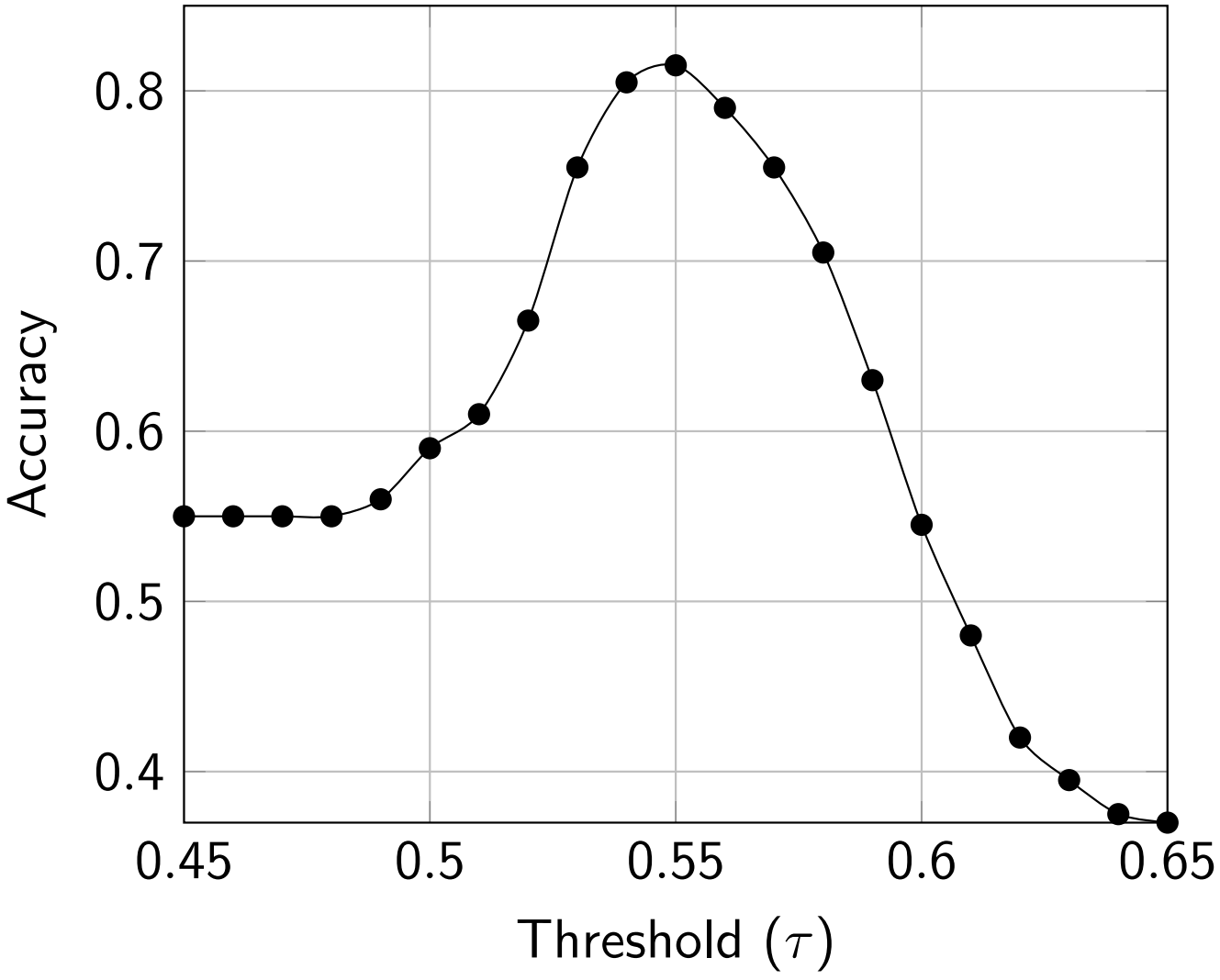
- Monge-Elkan for surface form similarity
- Word Mover Distance for content similarity

Imposed hard threshold (τ) on the final score

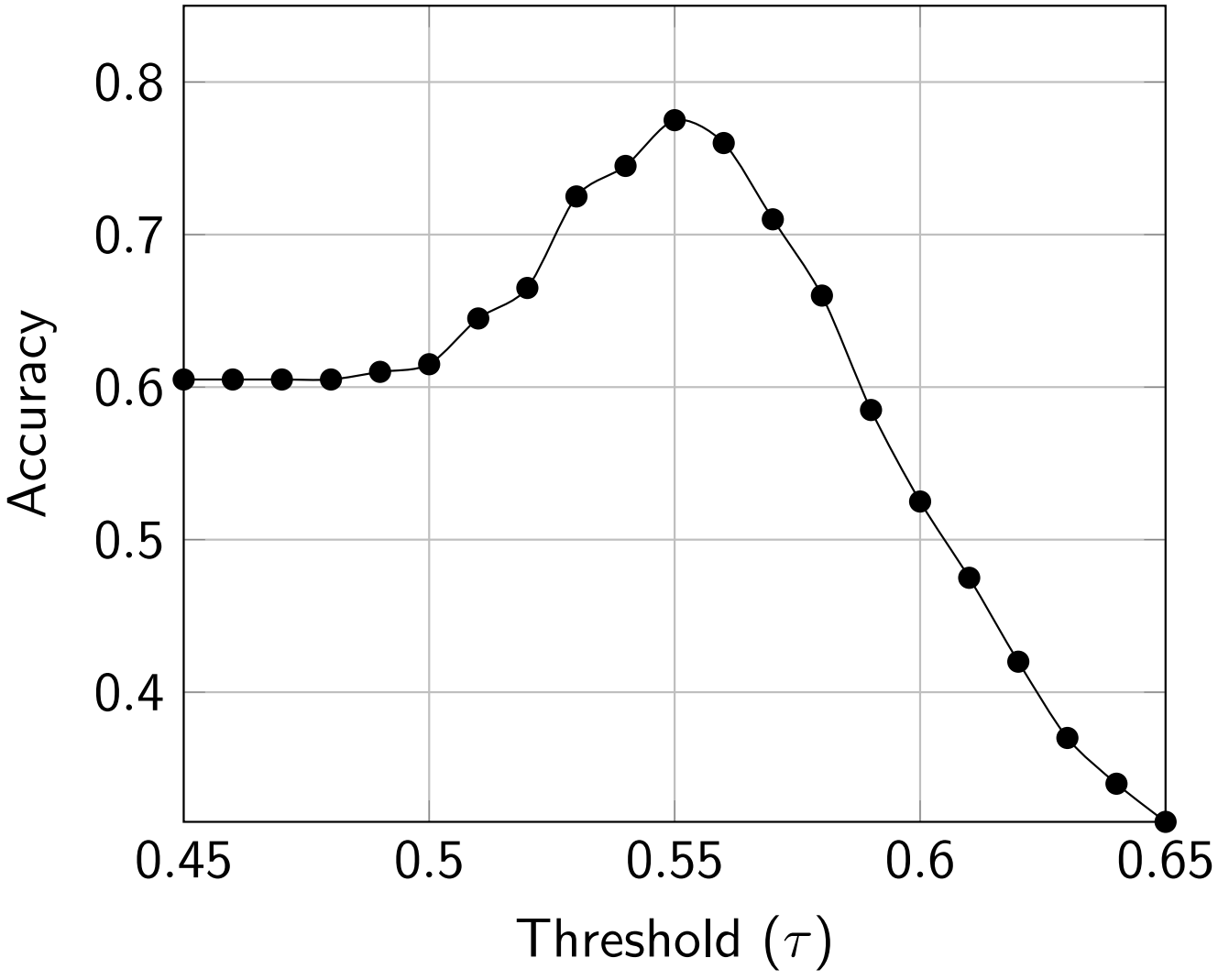
Evaluating DBpedia Mappings

- Evaluated the quality of mappings on a gold standard subset of 200 biographies from DIB and 200 biographies from ODNB.
- Problem is essentially an EL task, so evaluated with BAT framework.

Evaluating DBpedia Mappings (DIB)



Evaluating DBpedia Mappings (ODNB)



How Does It Affect Entity Linking?

Still testing, but the results look good.

Linking accuracy increased from 0.59 F1 to 0.73 F1 in GERBIL evaluation with Knowledge Base comprised of DBpedia, Geonames and DIB.

How Does It Affect Entity Linking?

Still testing, but the results look good.

Linking accuracy increased from 0.59 F1 to 0.73 F1 in GERBIL evaluation with Knowledge Base comprised of DBpedia, Geonames and DIB.

Again, still testing!

Thank You

 munnellg@tcd.ie

 github.com/munnellg