SocialLink

Linking DBpedia Entities to Corresponding Twitter Accounts

Yaroslav Nechaev (speaker), Francesco Corcoglioniti, Claudio Giuliano

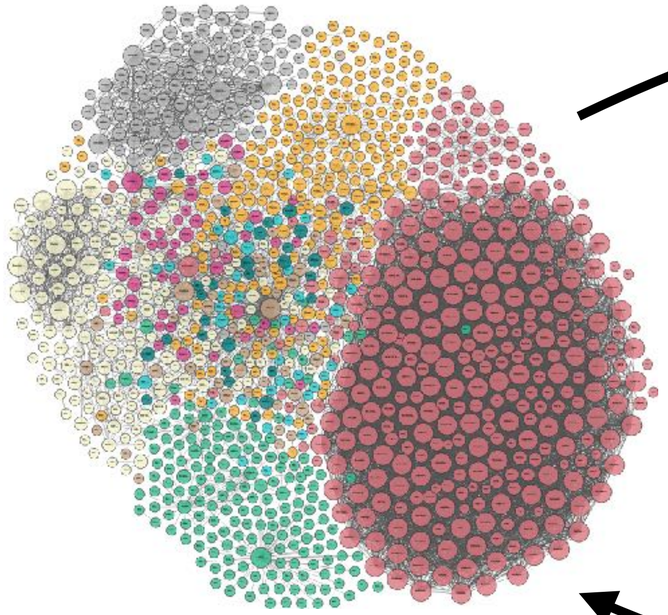
Linking Knowledge Bases to Social Media Profiles

- The goal is to bridge Linked Open Data cloud and social media

Linking Knowledge Bases to Social Media Profiles

- Trying to get the best of both worlds:
  - Knowledge Bases provide high-quality, structured, easily accessible knowledge
    - Cover wide range of entity types
    - Can contain obsolete knowledge
  - Social media is a vibrant source of up-to-date knowledge with an unparalleled coverage
    - More than 2B users in Facebook alone
    - Mainly covers living people, organisations and brands
    - Mostly unstructured

Use Cases



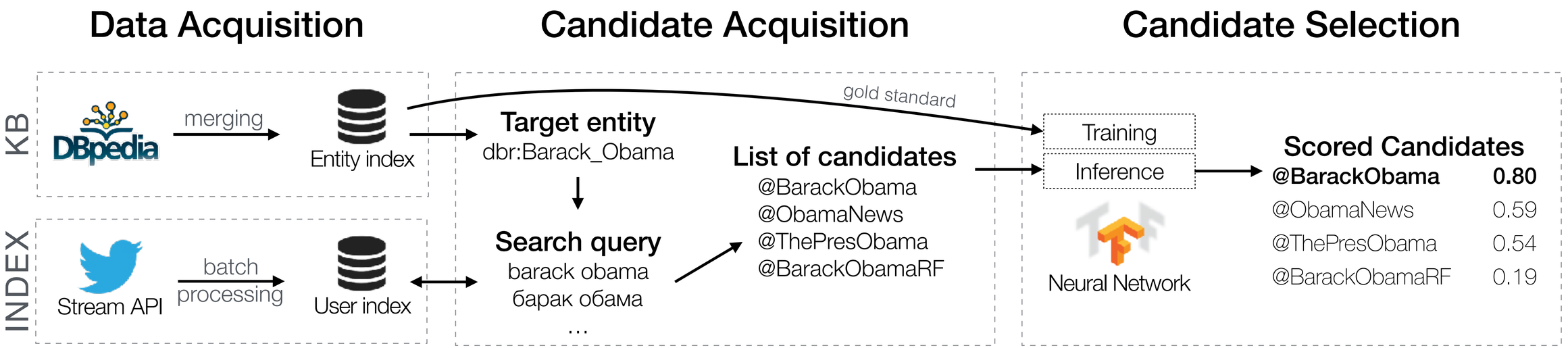
User profiling, entity linking, event detection, community detection and labelling

Enrichment, validation, referencing (providing sources to claims in KB)

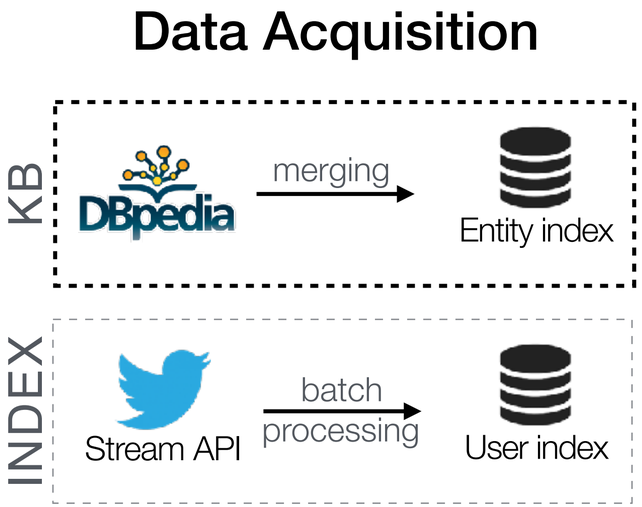
SocialLink

- Aligns people and organisations from 120 DBpedia language chapters to Twitter
- Considers 2.5 million entities
- Proposes candidates for 906k entities and aligns 271k entities
- The code and the resource are publicly accessible and updated
  - http://w3id.org/sociallink/
  - https://github.com/Remper/sociallink
  - Released in RDF, JSON and CSV, accessible through SPARQL endpoint

SocialLink Pipeline



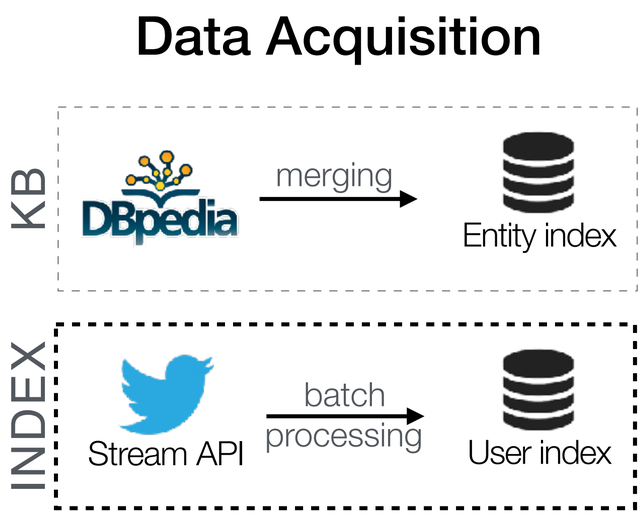
SocialLink Pipeline



- RDFpro[1] tool is used to download 120 DBpedia chapters and merge them along `owl:sameAs` links
  - Resulting Entity Index contains 1.4B triples
  - Considers living people (2M) and currently existing organisations (550k)
  - 58,745 gold standard alignments were extracted from `foaf:isPrimaryTopicOf` and `wikidata:P2002` properties

[1] Corcoglioniti, F., Rospocher, M., Mostarda, M., Amadori, M.: Processing billions of RDF triples on a single machine using streaming and sorting. In: ACM SAC. pp. 368–375 (2015)

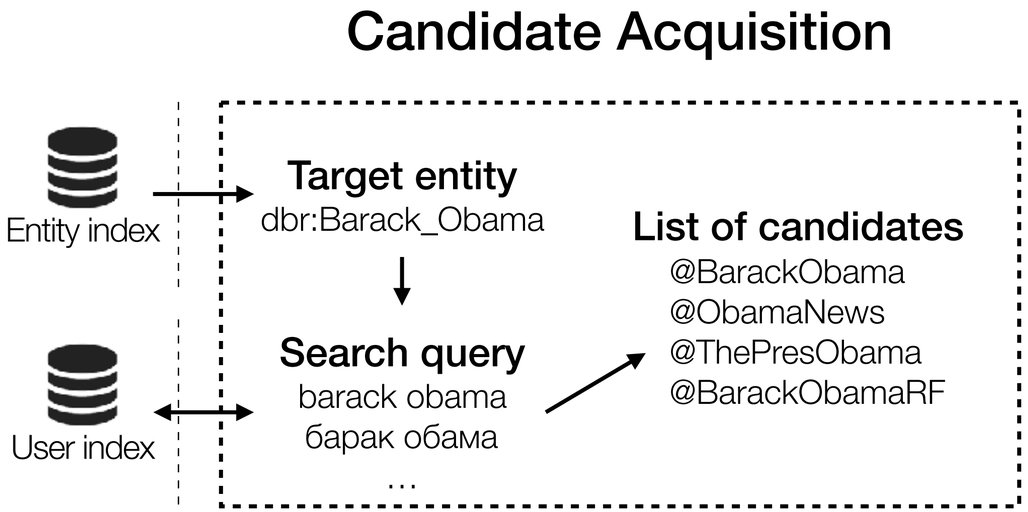
SocialLink Pipeline



- 3TB of Twitter data is processed and indexed using Apache Flink and stored in Postgres
  - Produced User Index contains 450GB worth of indexed data
  - Covers more than 240M Twitter users

[1] Corcoglioniti, F., Rospocher, M., Mostarda, M., Amadori, M.: Processing billions of RDF triples on a single machine using streaming and sorting. In: ACM SAC. pp. 368–375 (2015)

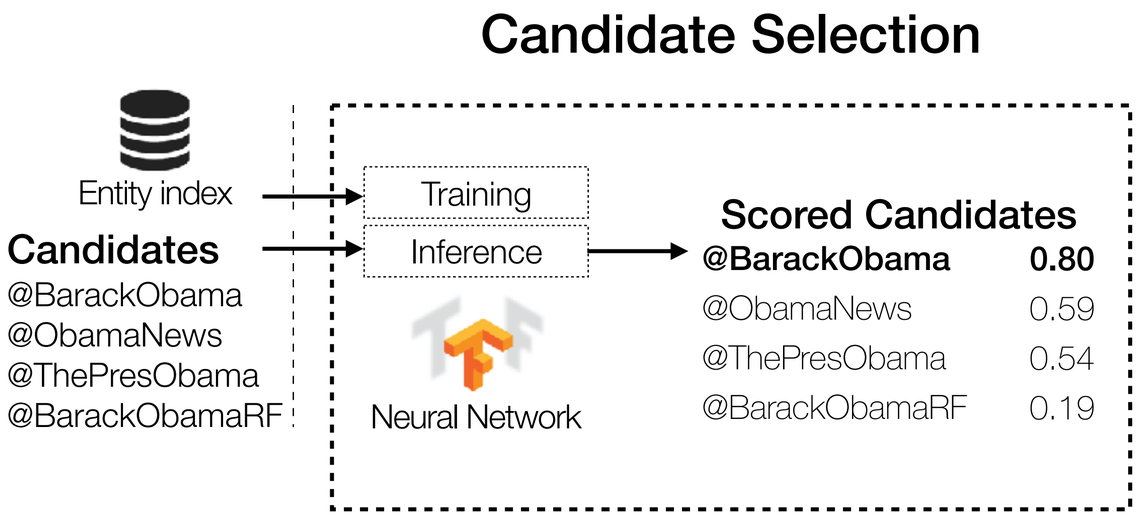
SocialLink Pipeline



- Produces a short list of candidate profiles for every entity in the Entity Index
  - 240M → ~300
- Name-based queries to User Index
  - Similar to our previous Twitter-based approach[2]
- Multiple queries for better recall

[2] Nechaev, Y., Corcoglioniti, F., Giuliano, C.: Linking knowledge bases to social media profiles. In: ACM SAC. pp. 145–150 (2017)

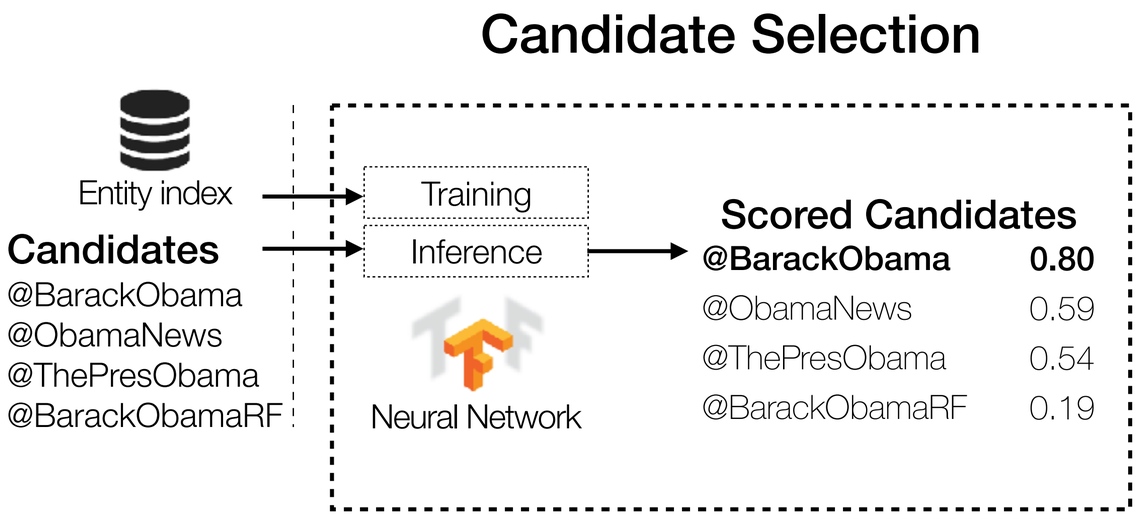
SocialLink Pipeline

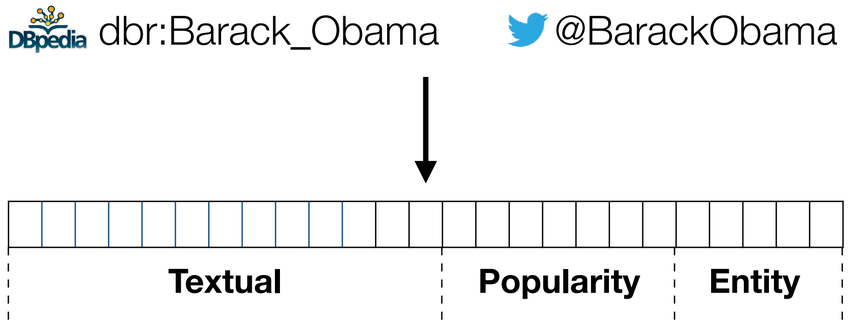


DBpedia dbr:Barack_Obama @BarackObama

[2] Nechaev, Y., Corcoglioniti, F., Giuliano, C.: Linking knowledge bases to social media profiles. In: ACM SAC. pp. 145–150 (2017)

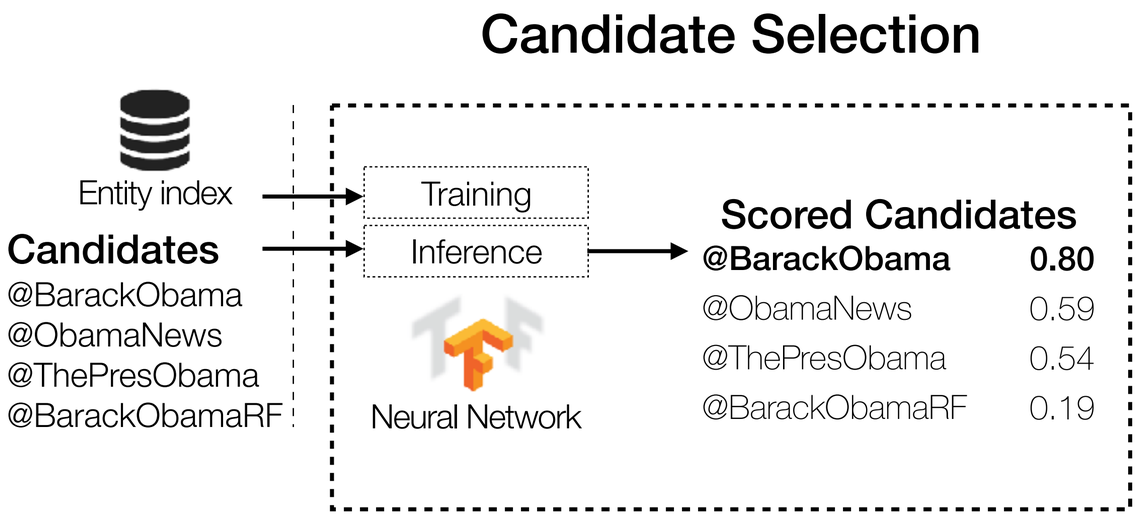
SocialLink Pipeline

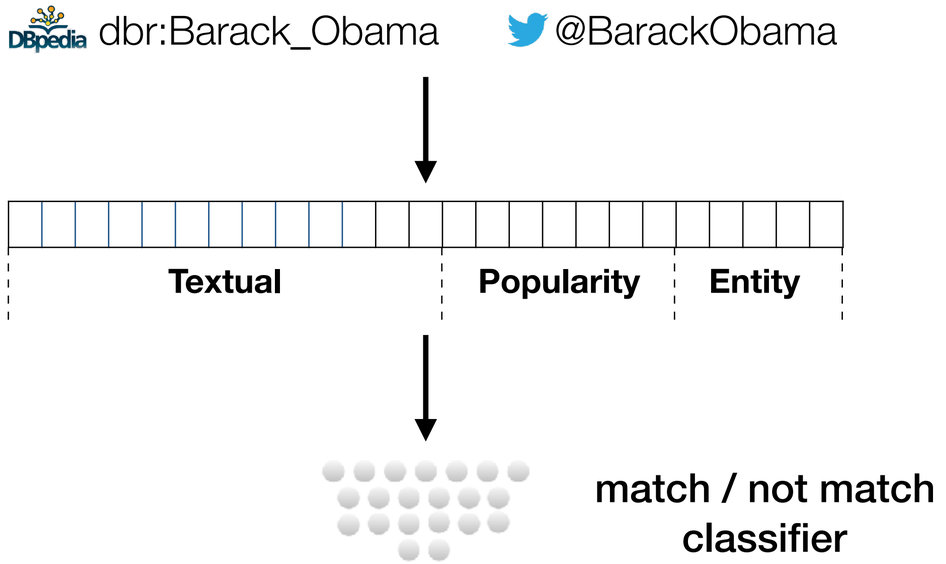




[2] Nechaev, Y., Corcoglioniti, F., Giuliano, C.: Linking knowledge bases to social media profiles. In: ACM SAC. pp. 145–150 (2017)

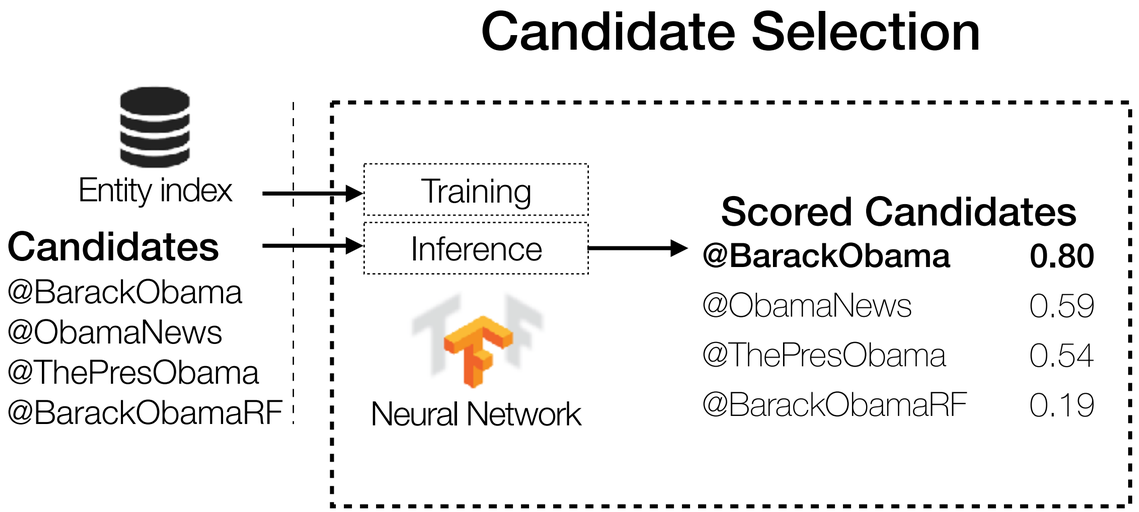
SocialLink Pipeline

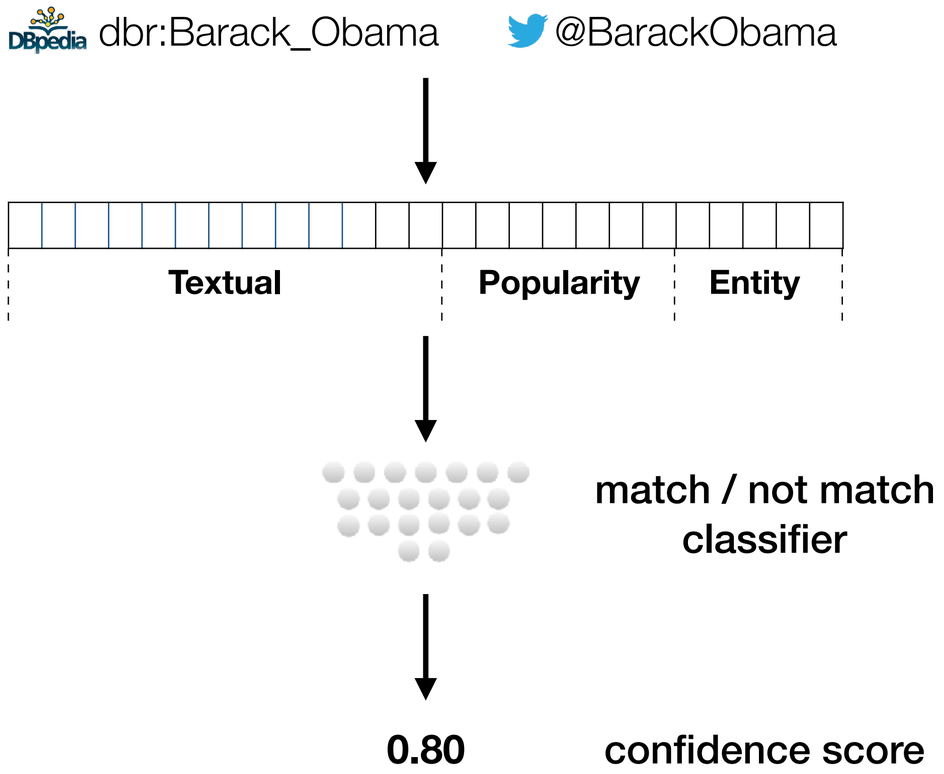




[2] Nechaev, Y., Corcoglioniti, F., Giuliano, C.: Linking knowledge bases to social media profiles. In: ACM SAC. pp. 145–150 (2017)

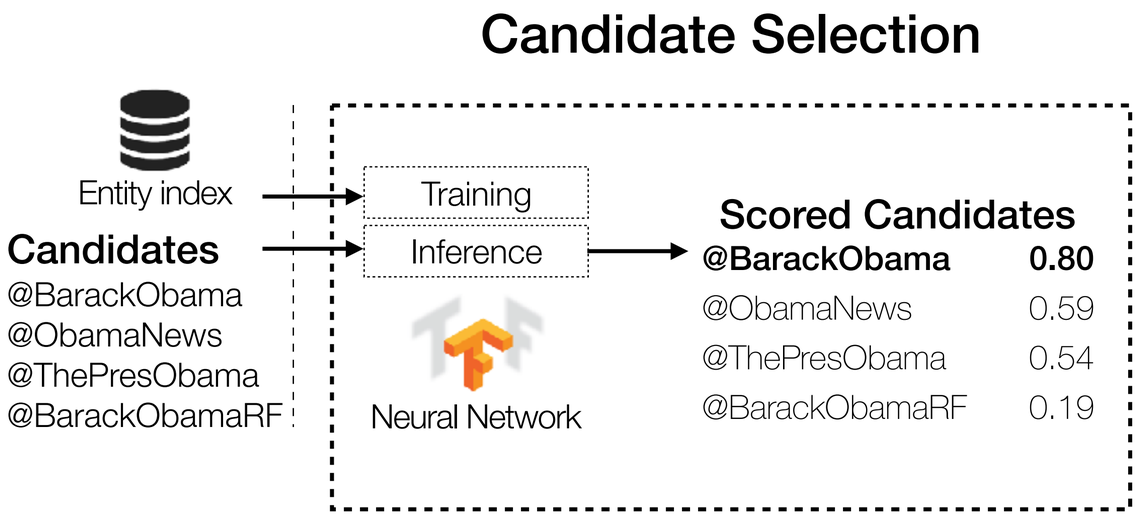
SocialLink Pipeline





[2] Nechaev, Y., Corcoglioniti, F., Giuliano, C.: Linking knowledge bases to social media profiles. In: ACM SAC. pp. 145–150 (2017)

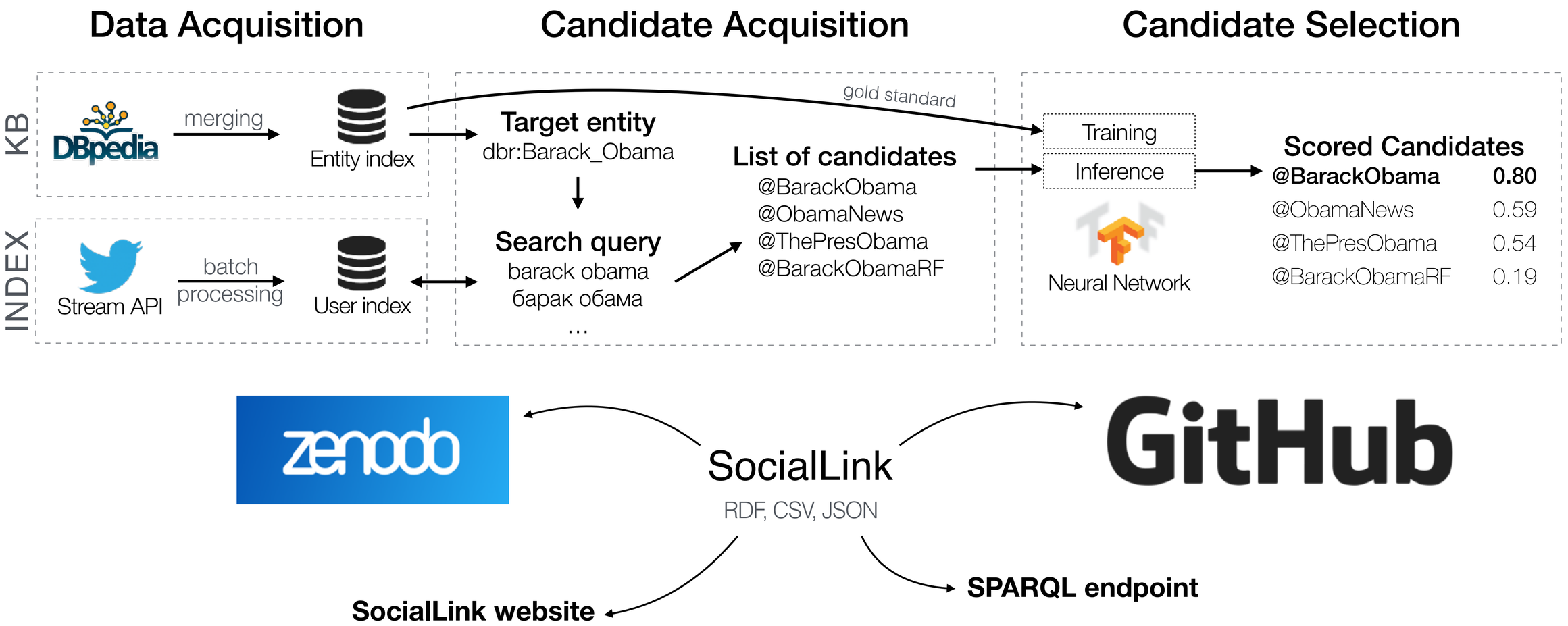
SocialLink Pipeline



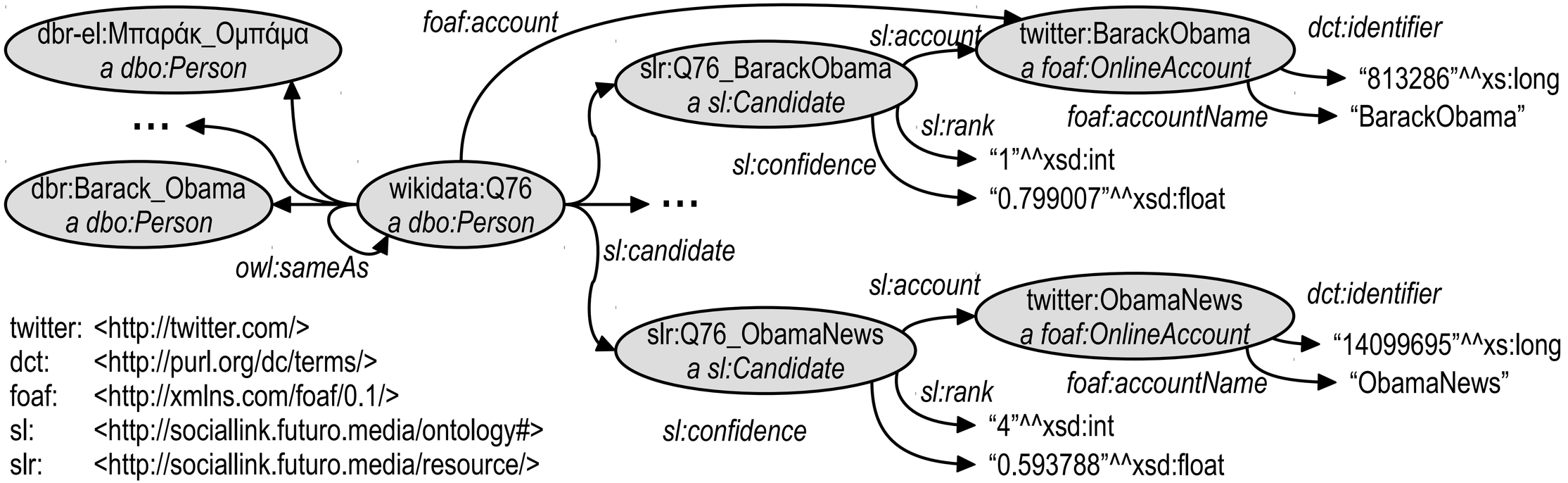
- Candidates are ranked based on the confidence score
- Alignment can be chosen aiming at the desired precision/recall balance
- Default thresholds exhibit 90% precision / 41% recall evaluated on the gold standard dataset

[2] Nechaev, Y., Corcoglioniti, F., Giuliano, C.: Linking knowledge bases to social media profiles. In: ACM SAC. pp. 145–150 (2017)

SocialLink Pipeline



RDF Representation




```
SELECT ?account {?e owl:sameAs <E>; foaf:onlineAccount ?account}
```

SocialLink Applications

- User profiling
- Entity linking
- Knowledge base enrichment

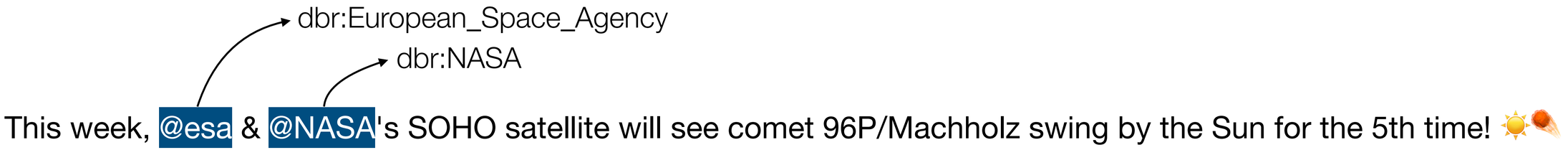
SocialLink for User Profiling

- Simplifies profiling pipelines.
  - Inference of user interests on Re-coding Black Mirror workshop[3]
  - Usage of linking was explored in other papers[4][5]
- Enables an automatic creation of gold standard datasets for the large spectrum of attributes (instead of manual annotation)
  - Paper in the works

[3] Nechaev, Y., Corcoglioniti, F., Giuliano, C.: Concealing Interests of Passive Users in Social Media. In: Re-coding Black Mirror workshop @ ISWC2017
[4] Besel, C., Schlotterer, J., Granitzer, M.: Inferring semantic interest profiles from Twitter followees: Does Twitter know better than your friends? In: ACM SAC2016
[5] Piao, G., Breslin, J.G.: Inferring user interests for passive users on Twitter by leveraging followee biographies. In: ECIR2017

SocialLink for Entity Linking

- Enables named entity linking to social media profiles
  - Implemented as part of the Social Media Toolkit[6]
- Directly disambiguates profile mentions in tweets against DBpedia
  - Used by second[6] and third[7] place entity linking systems in EVALITA2016 challenge



[6] Social Media Toolkit. https://github.com/Remper/sociallink/wiki/SMT-API
[7] Corcoglioniti, F., Palmero Aprosio, A., Nechaev, Y., Giuliano, C.: MicroNeel: Combining NLP tools to perform named entity detection and linking on microposts. In: EVALITA (2016)
[8] Minard, A., Qwaider, M.R.H., Magnini, B.: FBK-NLP at NEEL-IT: Active learning for domain adaptation. In: EVALITA (2016)

SocialLink for Knowledge Base Enrichment

- Direct import of up-to-date data to the knowledge base
- Validation and referencing of the existing data

- Proposed to Wikimedia community as a project – **soweego**[9]

[9] https://meta.wikimedia.org/wiki/Grants:Project/Hjfocs/soweego

Grants:Project/Hjfocs/soweego

< Grants:Project

Main page
Wikimedia News
Translations
Recent changes
Random page
Help
Babel

Community
Wikimedia Forum
Mailing lists
Requests
Babylon
Reports
Research
Planet Wikimedia

Beyond the Web
Meet Wikimedians
Events
Movement affiliates
Donate

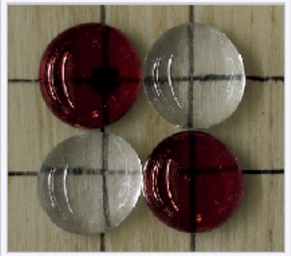
Print/export
Create a book
Download as PDF
Printable version

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Link by ID

In other languages

Project idea [edit]

soweego (**so***lid catalogs and* **wee***kee* **go** *together*) is a fully automatic robot that links existing Wikidata items about people to a set of reliable external catalogs.



soweego aligns different identifiers, but has nothing to do with the Go game!

Why: the problem [edit]

Data quality in a broad knowledge base (KB) like Wikidata is a vital aspect to secure confidence on its content, thus encouraging the development of effective ways to consume it. Wikidata is not meant to tell us the absolute truth: instead, we can see it as a container for different points of view (read *claims*) about every little fact of our world. These claims should always be verified against at least one reference to a trusted external source. However, this still does not seem to be the case. Despite specific efforts such as *WikiFactMine*[1] and *StrepHit*,[2] as well as more extensive community endeavors like *WikiCite*,[3] the lack of references is a critical issue that remains open. Recently, the problem was further highlighted by the Wikidata product manager in her presentation at WikiCite 2017,[4] where the following aspects about the KB statements clearly emerged:

- roughly **half** of them is totally **unreferenced**;
- less than **a quarter** of them has references to **non-wiki sources**;
- **most** reference values are **internal** links to other Wikidata **items**.

How: the solution [edit]

Alignment of Wikidata to structured databases can represent a complementary alternative to reference mining from unstructured data.[1][2] Think of an identifier as a reference to a Wikidata item: we can give force to the trustworthiness of that item if we manage to match it with the corresponding entry of an authoritative

status **proposed**

Project Grants

soweego

summary

soweego automatically links existing Wikidata items about people to *trusted* identifier catalogs. With **soweego**, our beloved knowledge base gets in sync with a giant volume of external information, and gets ready to become the universal linking hub of open data.

target

Wikidata

amount

75k EUR

grantee
- Hjfocs

advisor
- Magnus Manske

contact
- fossati@spaziodati.eu

volunteer
- TheDragonFire

this project needs...

volunteer

give feedback endorse join

Future Work

- Extension to Facebook and Instagram
  - Consistency checking across social media to improve results overall
- Candidate selection algorithm improvements
  - Better feature space using user and entity embeddings
  - Learning to rank instead of classification
  - Enable alignment to multiple accounts

Conclusions

- Introduced a resource for aligning DBpedia entities to Twitter profiles
- Applications for social media analysis and knowledge base enrichment
- Will be supported with new releases including:
  - Recent social media data
  - Pipeline and algorithm improvements

FONDAZIONE
BRUNO KESSLER

Thanks!

http://w3id.org/sociallink/

https://github.com/Remper/sociallink

Suggestions and/or pull requests are always welcome!