Named Entity Recognition on Open Data CSVs

Johann Mitloehner

Vienna University of Economics/Dep. Information Systems

February 2020

Open Data

- Governmental Open Data portals: rich source of free data
- Large part published as CSV
- 106,154 files from 977 portals downloaded by crawler
- Corpus will be made available

Portal	Files
cdn1.sdlabs.ru	41776
opendata.socrata.com	3835
webarchive.nationalarchives.gov.uk	2707
www.cps.gov.uk	1911
www.statweb.provincia.tn.it	1699
www.gov.uk	1645
daten.transparenz.hamburg.de	1247
dataservices.open.glasgow.gov.uk	1167
www-genesis.destatis.de	1083
www.statcan.gc.ca	1081
...	

Named Entity Recognition

- assign a type to a piece of text, e.g. Location, Person, Organization
- usually within context, such as one or more sentences

Machine learning approaches:

- context provides information on the likely tags
- less suitable for short pieces of text out of context
- such as values in database tables

Gazetteer:

- Directory of labels and types
- look up the value and assign all types for entry
- often results in
  - no type
  - set of unrelated types
    type of column = intersection of type sets for all values
- Open Source: crowd effort, varying coverage and quality

DBpedia and Wikidata

- DBpedia: extract structured data from Wikipedia infoboxes
- Wikidata: free knowledge base that anyone can edit

  <http://dbpedia.org/resource/Abraham_Lincoln>
  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
  <http://dbpedia.org/ontology/Person> .

File	Statements
instance_types_transitive_en.ttl	31,254,272
instance_types_transitive_de.ttl	9,669,523
redirects_en.ttl	7,699,478
redirects_de.ttl	1,356,266
wikidata_types.ttl	22,980,019
wikidata_en_de_labels.ttl	21,014,641
wikidata_subclasses.ttl	1,754,163
total	95,728,362

- Wikidata labels in all languages: 128,693,935
- DBPedia: transitive types for all entities available as dump files
- Wikidata: transitive types from Virtuoso Sparql query
  import and query about 16 hours

From Knowledge Graph to Gazetteer

Entities and Labels

Abraham Lincoln, 16th president of the United States:

dbpedia.org/ontology/Person, xmlns.com/foaf/0.1/Person, www.w3.org/2002/07/owl#Thing, schema.org/Person, dbpedia.org/ontology/Agent ..

but also ships, high schools, bridges named Abraham Lincoln

- number of types for label 'Abraham Lincoln' much larger.
- other labels for entity: Abe Lincoln, Honest Abe, President Lincoln..
- huge number of possible types

Multi-linguality

- only English and German, all Wikidata and DBpedia languages not feasible with our resources

Numeric Values

- many numbers with entries, such as historically significant years
- hard to automatically identify NE vs numeric value
- ignore numeric values

CSV Table Processing

- corpus contains small number of very large CSV files
- impact performance and feasibility with given resources
- assumed that
  - reference tables tend to be small or moderate size
  - references from huge tables not contribute much to results

Only files smaller than 500k were used in the analysis

- resulting in about 95,000 files with a total size of about 3.5 GB
- about 10,000 files bigger than 500k totalling 170 GB were excluded
- Corpus is downloaded by crawler with post-processing
- number of read errors low

Error	Number
header only	605
no rows in table	1

Number of Tables

Tables	Header	Columns	Avg Rows	Avg Cols
95,121	83,052	1,916,560	273.5	20.1

Rows	Tables	Header	Avg Cols
10	33716	0.92	28.9
20	11534	0.83	13.7
30	6138	0.84	15.3
40	4272	0.85	14.9
50	4766	0.90	16.9
65447	34695	0.85	15.7

- Columns that contain only numbers or null are ignored
- Tables that are left without columns are dropped

Tables	Columns	Avg Cols
92,409	1,110,321	12.0

Most Frequent Values

Value	N_V	N_T	Types
ШТ	1124410	0	
Department of Health	484258	5	wd/health ministry, wd/department of the United Kingdom Government, wd/government agency
УПАК	308298	0	
tonnes	287576	4	dbpedia.org/ontology/Work, schema.org/-MusicAlbum, dbpedia.org/ontology/-MusicalWork, schema.org/CreativeWork
КГ	278198	0	
F2F	214898	0	
Glasgow City	207396	22	wd/constituency, dbpedia.org/ontology/-Place, wd/proper noun, wd/painting
Scotland	205939	27	wd/constituency, wd/article, wd/series, wd/constituency of the Parliament of Great Britain, wd/group
D	189279	118	wd/Philippine TV series, wd/serial, wd/rail transport, wd/generic programming language, wd/Wikimedia list article
$\star$	189144	1	wd/character

Most frequent Types

Type	Number
dbpedia.org/ontology/Location	9584
schema.org/Place	9584
wd/entity	8628
wd/group	8143
dbpedia.org/ontology/Place	7604
wd/Wikibase item	6854
wd/Wikidata property	6854
wd/intellectual work	6541
wd/Wikidata property related to economics	6534
wd/unit of length	6534
wd/English units of measurement	6529
dbpedia.org/ontology/PopulatedPlace	6489
dbpedia.org/ontology/Agent	6349
www.ontologydesignpatterns.org/ont/dul/DUL.owl#Agent	6349
schema.org/Organization	6090

Values and Types

- About 19% of values associated with at least one type

	Total	With Type	Percent
Values	94,347,545	17,528,442	0.186
Unique	10,374,757	234,644	0.023

- Selectivity: number of distinct values divided by the number of values
- Columns with at least three values:

Columns	Avg Number of Values	Avg Distinct Values	Selectivity
458,333	204.3	51.1	0.2

Type Coverage

Fraction of type covered in column
E.g. coverage = 1 for a column with names of all 50 US states

Type coverage for columns with

- at least three distinct values and
- a fraction of 0.8 of typed values

Columns	Avg Coverage
72149	0.017

- a little better for selectivity 1:

Columns	Avg Coverage
6430	0.019

- sadly, down again for fraction 1:

Columns	Avg Coverage
1477	0.012

- complete coverage of at least one type, for any frac, sel

Columns
40

Types and Reference Tables

Direct approach: compare values in all columns in all tables

Type information for identifying reference tables

- how many parent and child values share the same type?
- can the search be sped up with type information?

Define Reference:

- Strict: all values in child must be present in parent
- Loose: 90% of values in child must be present in parent
- Sets of values rather than the original lists

Idea: look for parent values in the first column of the parent table

- not feasible, would result in too many misses.

Index	Tables
0	38065
1	34797
3	15861
2	11139
5	5732
4	5718
6	4332

Restrictions on Reference Columns

To limit the amount of processing:

- The number of distinct values must be at least 10
- The selectivity must be 1

Candidate Tables
22480

- Still need to check with each column of every other table
- Each check computes intersection of the two column value sets

Various other approaches were tried:

- use statistics to eliminate columns, e.g. sort order
- use compiled Cython module for the check

However

- the number of values is very small in most cases
- plain Python set intersection is the fastest method

Loose vs Strict Reference

Loose check identifies about 15,000 tables i.e. 16% possible reference tables

Number of Reference Tables
15054

Perhaps surprisingly, strict check decreases only slightly to 15%.

Number of Reference Tables (strict)
14290

- most columns contain a very small number of values
- further reduced by taking sets from lists
- Total run time about 4 hours on 8x Xeon E3, 100 GB RAM
- Python library `ray` was used for parallel processing
  - arbitrary objects in shared memory
  - github.com/ray-project/ray

Using Type Information for Finding Reference Tables

Idea: If two columns share one or more common type for all values then they are likely candidates for a referencing relationship.

- Can the type information provide a shortcut to finding reference tables?
- Unfortunately, at least in this data set, most reference tables would by missed by taking that shortcut: [1]

Common Type	Tables
0	14768
1	1005

- Only about 6% of the reference tables also share at least one common type in at least one column.

[1] Note that the values do not add up to the number of reference tables in the previous section since there can be several references among two tables, with or without common types in the corresponding columns.

Conclusions and Future Work

- Type data no use to improve performance in finding references
- Subset check succeeding does not imply a referencing relationship
- The number of actual references is probably much lower
- Determine actual number by manual examination
  - How many subsets are actual references?
  - How does the type data correspond to results of manual checking?

→ Manually find references for a sufficiently large part of the corpus

- Evaluate the NE approach
- Test machine learning methods