Distant supervision for Numerical Relation Extraction

Knowledge Base

- Derived from `data.worldbank.org`, 4371979 numerical facts about 249 countries, 1281 attributes

/m/04g5k	3126000130	EG.ELC.PROD.KH
/m/02k8k	1969.179	EN.ATM.CO2E.KT
/m/06nnj	332315	SP.POP.TOTL
/m/019rg5	55.020073	SP.DYN.LE00.IN
/m/05sb1	19974.148	EN.ATM.CO2E.KT
/m/05v8c	10000000000	EG.ELC.PROD.KH
/m/03spz	7639000100	EG.ELC.PROD.KH
/m/06vbd	44249.688	EN.ATM.CO2E.KT
/m/0d060g	51.3	IT.NET.USER.P2
/m/05qkp	62.298927	SP.DYN.LE00.IN

Selected Relations

Relation Name	Relation Code
Land area (sq. km)	AG.LND.TOTL.K2
Foreign direct investment, net (current US$)	BN.KLT.DINV.CD
Goods exports (current US$)	BX.GSR.MRCH.CD
Electricity production (kWh)	EG.ELC.PROD.KH
CO2 emissions (kt)	EN.ATM.CO2E.KT
Pump price for diesel fuel (US$ per liter)	EP.PMP.DESL.CD
Inflation, consumer prices (annual %)	FP.CPI.TOTL.ZG
Internet users (per 100 people)	IT.NET.USER.P2
GDP (current US$)	NY.GDP.MKTP.CD
Life expectancy at birth, total (years)	SP.DYN.LE00.IN
Population (Total)	SP.POP.TOTL

Corpus

- Subset of the tac corpus
- 268, 036 Documents, X sentences having a country and a number
- List of countries augmented manually by adding all possible synonyms (Dutch, Netherlands) and inflections (Ireland, Irish)

Distant supervision process for numerical relation extraction

- Extract a country and number from a sentence, go to kb and check if there is a match.
- Can be creative during matching:
  - Distance based matching
  - Time based matching

False Positives

Numbers are weak entities

- Vanilla numerical relation matching is bound to attract humoungous amounts of false positives;
- Stems from the fact that numbers don't have an identity of their own.
- Consider India and Mumbai Vs. India and 19
- Mumbai is a strong entity, 19 is a **weak** entity.

False Positives

Numbers are weak entities

- Compare the number of sentences in which India and Mumbai appear together, vs the number of sentences in which India and 19 appear together.
- Just 19 can be appear with India in several contexts:
  - Internet user %
  - Billion dollars invested by a company
  - % of people below the poverty line
  - date (if we are not careful), number of medals won by Indian athletes...
- Can we expect the situation to be even worse for certain types of numbers?