sg

Occurrence Statistics of Entities, Relations and Types on the Web

Submitted in partial fulfillment of requirements for the degree of Master of Technology

by

Aman Madaan

Under the Guidance of Prof. Sunita Sarawagi

Department of Computer Science and Engineering
Indian Institute of Technology Bombay

April, 2014

Occurrence Statistics of Entities, Relations and Types on the Web

October 21, 2014

Abstract

The task of numerical relation extraction poses new, hitherto untackled challenges. The bewildering amount of false positives, units, modifiers, time varying relations are just some of the issues that are non existent for standard relation extraction, but become crucial when numbers are involved.

We discuss molding distant supervision for numerical relation extraction. The standard one to one mapping using numbers as the second entity fails. Adding units help in improving the training data to a certain extent. The third heuristic, keyword based pruning yields drastic improvement in training data. The analysis of the results from the three heuristics lead to a basic rule based extractor, which performs better than any of the heuristics.

1 Introduction

Massive knowledge bases containing the entire information of the web in a neat, ready to process, structured form continue to be a part of an IR researcher's reverie. This is not unexpected; such knowledge bases have potential of revolutionizing the way in which information is searched by users on the web, or exchanged by machines among themselves. Clearly, any progress towards a solution will have to deal with intricacies of how facts are expressed in the natural language. It turns out, that such intricacies are too many to exhaust. Years of research has went into the aforementioned goal, and still are plenty of loose screws. This report discusses one of them, numerical relation extraction.

Section 2 defines the problem. Section 3 lays the foundation for distant supervision.

1.1 Terminology

This report will demand that the reader is on the same page vis-à-vis a few terminologies.

- **a)Entity** An entity is something that exists in itself, actually or hypothetically. [6]

- **b)Relation** A relation specifies a concept which binds two entities. For example, creator(Linux, Linus Torvalds).
- **c)Mention** A piece of text which expresses a relation. For example, the sentence "Linus Torvalds is the creator of Linux."
- **d)Match** A mention expressing a relation R is called a match for R. The criteria for deciding whether the mention m is a match or not can vary. Presence of both the entities is a however a mandatory condition.
- **e)Extraction** A 3-tuple $R(A, B)$ where A and B are entities related via R.

2 Problem

Train extractors that can harness the Web for numerical relations, where relations are 3-tuples linking an entity to a number. For example,

- (India, **economy**, 1.842 trillion USD)
- (China, **internet users**, 590.56 million)
- (USA, **land area**, 2,959,054 square mile)

3 Snowballing

Let us motivate the idea by considering the following related problem:

Suppose we want to populate the repository of founders of companies, and all that we know for a fact is that Elon Musk is the founder of SpaceX. The problem can be divided into two parts, each of them rely on an intuition about how human beings form sentences in general.

- The first of them is given an entity pair, and a corpus of documents, find out all the sentences that express a relation between the entity-pair.

  Command line ninjas will quickly think of the following solution: grep -i 'entity1' sentences—grep -i 'entity2'

  The intuition behind this perhaps the most obvious solution is that *a sentence that houses both the entities can be expected to express a relation between them.* A quick web search with the query "entity1" and "entity2" will show that this intuition is not out of the blue.

- Sentence structure depends on the relation being expressed. In verbose, if two sentences express the same relation, there will be (okay, there can be expected to be) many *features* that are similar in both of them. These include POS tags, words around the entities, dependency path between the entities to name a few.

Putting together the intuitions above, we can solve the problem as follows: Collect all the sentences which have SpaceX and Elon Musk in them, extract features from these sentences. Favor those features which repeat. Now us this set of features to extract similar pairs from other sentences. A fancier solution

would be to re use the extractions to learn more features, and continuing the process till the point of diminishing returns.

This seemingly shaky method actually works [1] and is popular by the name of snowballing.

4 Distant Supervision: Snowball scaled up

4.1 Introduction

If the idea of Snowball looks convincing, Distant supervision should follow naturally. Earlier, we considered only one relation for one entity pair. Scale the amount of both of these up and we have distant supervision. The basic setup is as follows:

- a) KB : A knowledge base consisting of facts. The facts are 3-tuples; the entities and the corresponding relation. For example:

Entity	Entity	Relation
Donald Knuth	Wisconsin	Born In
Srinivasa Ramanujan	Erode	Born In
Alan Turing	London	Born In
Alon Musk	SpaceX	Founder Of

  has 4 different facts

- b) Corpus The repository of text where we expect to find the sentences that express facts that we know. We need another repository, called the test set, where we will run our extractor to obtain new facts. These two can be the same.

4.2 Distant Supervision Assumption

Every sentence that has an entity pair (e_1, e_2) expresses the relation which exists between (e_1, e_2).

4.3 Matching

We next need to align our knowledge base with the corpus. This process is also called matching.

4.4 Training

Recall that obtaining the sentences which express a relation gives us training data, which we want to use to learn relation extractors, our goal. There are several ways to achieve this, starting from the naive ways of training sentence level classifier extending to fancier graphical model based learning.

We briefly discuss the different training methods as we look at a survey of works on Distant Supervision so far.

Data: Corpus C, Knowledge Base KB
Result: Training data, D, A set of matches
Break C into a set of sentences, S;
for *each sentence s in S* **do**
  let E = all entity pairs in s;
  for *each entity pair (e_1, e_2) in E* **do**
    if ∃ *relation r in KB with r(e_1, e_2)* **then**
      add s to D with label r
    end
  end
end

Algorithm 1: Distant Supervision

5 Survey

The first distant supervision paper came out in 1999. Since then, almost every knob that could be twisted in the ds machinery, has been twisted. Different types of relations and different types of datasets will pose different challenges of course, and this report deals with some of them.

5.1 Constructing Biological Knowledge Bases by Extracting Information from Text Sources, Craven and Kumlien 1999

The was the first work to use distant supervision for creating a repository of biological facts. They targeted 5 different relations between Proteins, Tissues, Cell-Types, Diseases, Pharmacologic-Agents and Subcellular-structure. A naive bayes based simple relation extractor is first described. The extractor works in 2 steps: A classifier trained on hand labeled data first labels whether a sentence *can* express a particular relation. If yes, the sentence is searched for a pair of one of the 5 types depending on the label and the fact is added to the database. The authors note that it took an expert 35 hours to hand label the corpus. This forms the basis of motivating distant supervision based methods.The paper also identifies many areas of improvements, like a pair of entities taking up multiple relations, on which subsequent work has been built up.

The authors obtain an improvement of around 9 precision points with distant supervision. It is possible that the following points contributed to the improvement in scores:

- Constrained entity set. Proteins and Tissues won't just appear together without any possible relation.
- The corpus used was very well aligned with the knowledge base.

5.2 Distant Supervision for the Web, Mintz et. al 2009

This work revived the interest of the community around the problem of distant supervision. The major contributions of this paper can be listed as follows:

- **Distant Supervision for the Web** Craven and Kumlien had a limited knowledge base and a limited corpus to align it with. As the web exploded, the knowledge bases that became available were of the magnitude of FreeBase, and the corpus that can be aligned with it was the web. Mintz et. al bought this fact to the spotlight and sparked a series of works.
- **Features** They designed a set of rich features that are used by researchers to date. These features include dependency paths, POS tag sequences, word sequences to name a few.

5.3 Learning 5000 relation extractors

The number of relations typically used in the works were limited to the range of 30-40. This work targeted 5000 relations, which is a big jump. To fight sparsity that will arise in the training data due to a large number of classes, they also train a top level classifier which decides which extractor should be used for a particular document. Not all the extractors are employed for a given document.

5.4 Riedel et. al

Distant supervision assumption does not really holds when the knowledge base is not well aligned with the corpus. Another way of saying this would be that the corpus can be expected to consist of sentences that can host entity pairs in a wide range of contexts, and the sentences may have nothing to do with the relations in which the entity pairs appear. They take the example of the relation "nationality". A wide range of popular entities that are born in the country will be mentioned in the sentences that may have nothing to do with the fact that they were born in that country. In such cases, it can be argued that the training data generated will be extremely noisy and will lead to poor extraction performance. Riedel et. al relax the distant supervision assumption.

5.5 MultiR

Some of the entity pairs can occur in multiple relationships. For example,

6 Numerical Relations

Numerical relations are much like the usual entity-entity relations, just more problematic. What we call a *numerical relation* will usually be called an attribute colloquially. For example, Camel and 215cm are related via the relation "Average Height".

7 Numerical Relation extraction using MultiR

The problem of numerical relation extraction can be easily moulded into the distant supervision framework. Recall from section [4] that we need a corpus and a knowledge base to generate the training data for distant supervision. We also need to fix on the set of relations that will be targeted.

7.1 Relations

We selected the following 10 relations for our experiments:

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

The relations were selected keeping in mind the availability of sentences which potentially express these facts along with difference in units.

7.2 Knowledge base

Our knowledge base was derived scrapped from `data.worldbank.org`. It has 4371979 numerical facts about 249 countries ranging over 1281 different attributes.

Country id	Number	Relation
/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

8 L'homme propose, et Dieu dispose

Our life would have been simpler if MultiR would do the desired, extract the numerical relations from the corpus provided information about some of the numerical relations. But as it happens, it was not meant to be. The model not doing well may mean several things. Low quality training data being one of the primary reasons. Indeed, it turned out that vanilla distant supervision leads to an **unprecedented** amount of false positives, which are the root cause of everything that went wrong in the distant supervision pipeline.

8.1 Numbers and False positives

Why would numbers lead to false positives in the first place? The problem stems from the fact that numbers have no identity of their own; they represent

count of some real entity or phenomenon.

- **Numbers can appear in many more contexts with an entity** The number of ways in which any two entities can appear together in a sentence is far less than the number of ways in which a number and a quantity can appear together. For example, Consider the entity pair "Bill Gates" and "Microsoft" and the entity-number pair "Bill Gates" and "3" (say). While former will usually co-occur in finite contexts (Founder, CEO, Evangelist etc.), the latter may co-occur anywhere Bill Gates happen to be around something which is 3, the number of cars, billion dollars donated, number of units headed, position in the company, number of business units shutdown by Microsoft and so on.

- The situation is worse for smaller whole numbers, which are more frequent. This intuitively makes sense as we are more often see 2,3 or 11 than 111212233 or 11.42143.

- **The match mines**

  During the initial phases of our experiments, we stumbled upon the *match mines* These were basically huge tables, world cup scores of all the matches played and so on. A couple of such sentences were responsible for 21% of the matches! It is easy (and very important) to get rid of such sentences. For subsequent runs, we first sort the sentences by length and then remove top 1000 of them.

The first point of debugging a model which is not performing as well as it should is doing sanity check on the training data.

9 Fighting False Positives with units

Analyzing results of plain matching made it clear that units will help in improving both precision (by eliminating matches where the unit is not present) and recall (increasing matches by canonicalization of numbers and conversion to SI units). We found that though units helped in drastically cutting down the number of false positives (match mines were completely eliminated), and helped recall (lots of good matches for Land area and Population), the number of false positives was still a trouble. The number of false positives was typically high for cases where the unit was percentage, since it is again a very generic unit. For other relations too, the number of false positives was very large. The large number of false positives, apart from degrading quality of the model, make evaluating the quality of matcher very difficult.

As the next step, we integrated Prof. Sunita's unit extractor in the distant supervision pipeline. Concretely, we fed the corpus to the unit extractor to get information of the following form (sample):

2000064	172010.0::64:72;500.0::157:160;
2000077	75.0:united states dollar:17:19;2010.0::55:67;29.0::69:7
2000080	50.0::47:49;
2000112	172009.0::56:64;
2000113	0.10000000149011612::58:61;
2000120	3.15564E7:second:113:115;1.0::131:132;

Where the format of a single match is sentence id [TAB] Number, unit, startOffset, endOffset;Number2,unit,startOff,endOff The knowledge base was also converted to SI units. We see lots of true positives for attributes that take value in the higher range. The number of false positives has also drastically decreased. Units have been particularly helpful in dealing with the so called "match mines", sentences which have lots of country names and numbers, like score tallies.

10 Numbers are weak entities: a case for keywords

10.1 Motivation

Matches from unit extraction showed that in some cases, the sentence that supposedly labeled as a match for a particular relation has no mention of the relation itself at all. For example, consider: "In eurozone powerhouse Germany, industrial orders jumped 3.2 percent in June, official data showed Thursday, with foreign demand behind a sharp rebound following a surprise drop in May." In this sentence, (Germany, 3.2) was considered as a match pair for the relation Internet user percent. Clearly, it has nothing to do with it.

10.2 Numerical Relations are Explicit

A key observation that can be made by going through the sentences that express numerical relations is that one cannot be too poetic while forming a sentence that is supposed to state a numerical fact. This is in stark contrast with sentences expressing relations between entity pairs, wherein the underlying relation might be implicit. If we want to state GDP of a country in a sentence, there is no escape from the words like "GPD" or "gross domestic product" and the likes * .

Compare this with a sentence that must relate Microsoft and Bill Gates. A few ways of stating that Microsoft was founded by Bill Gates can be enumerated as follows:

- Bill Gates is the founder of Microsoft
- Bill Gates founded Microsoft
- Bill Gates is the father of Microsoft
- Bill Gates laid the foundation stone of Microsoft
- Bill Gates started Microsoft

If this is indeed true, imposing an additional constraint of keyword being present in a sentence in addition to the fact being present can help in cutting down the number of false positives. We note that such a pruning is possible only in case of numerical relations. As mentioned earlier, for real world entity pairs, co-incidental matches will be rarer and a constraint on the relation word being present will be too restrictive.

10.3 Approach

Let M_r be the set of matches obtained by standard unit + distance based matching for a relation r. We prune M_r by picking only sentences which contain one of the words in the set keywords(r). The sets keywords(r) are manually crafted.

Relation Keywords	(case insensitive)
Internet User %	"Internet"
Land Area	"area", "land", "land area"
Population	"Population"
Diesel	"diesel"
GDP	"Gross domestic", "GDP"
CO2	"Carbon", "Carbon Emission", "CO2"
Inflation	"Inflation", "Price Rise"
FDI	"Foreign", "FDI"
Goods Export	"goods"
Life Expectancy	"life", "life expectancy"
Electricity Production	"Electricity

11 Results

Before discussing the results, let us summarize the heuristics:

a) Vanilla: A sentence having a country-number pair $c - n$ is called a match for a relation R if $R(c, n)$ is a fact in the KB. b) Unit Based: Feed the corpus to a unit extractor, extract normalized numbers along with units. During matching, both the normalized value of a number and the unit are considered. c) Keyword based: In addition to matching number and unit, also check whether one of the keywords for a relation is present.

11.1 Vanilla Matching: Results

Relation	Total Matches	Sampled Matches	True Matches	Precision(%)
Land Area	1884	15	**1**	**6.7**
Foreign Direct Investment	0	0	0	0
Goods Export	0	0	0	0
Electricity Production	381	10	0	0
CO_2 Emission	0	0	0	0
Diesel Prices	8491	15	0	0
Inflation(%)	8689	15	0	0
Internet Users(%)	182319	40	0	0
GDP($)	0	0	0	0
Life Expectancy	267	10	0	0
Total Population	0	0	0	0

11.2 Units Based

Relation	Total Matches	Sampled Matches	True Matches	Precision(%)
Land Area	98	40	32	80
Foreign Direct Investment	791	40	1	2.5
Goods Export	816	40	3	7.5
Electricity Production	19	19	0	0
CO_2 Emission	196	40	2	5
Diesel Prices	2	2	2	100
Inflation(%)	27598	40	0	0
Internet Users(%)	24639	40	0	0
GDP($)	1790	40	0	0
Life Expectancy	3081	40	0	0
Total Population	5225	40	11	27.5

Table 1: Unit Based

Relation	Total Matches	Sampled Matches	True Matches	Precision(%)
Land Area	61	40	**30**	**75**
Foreign Direct Investment	8	0	0	0
Goods Export	4	4	0	0
Electricity Production	0	0	0	0
CO_2 Emission	16	16	2	12.5
Diesel Prices	2	2	2	100
Inflation(%)	3853	40	25	62.5
Internet Users(%)	308	40	7	17.5
GDP($)	0	0	0	0
Life Expectancy	99	40	15	37.5
Total Population	607	40	24	60

Table 2: Keyword Based

11.3 Keyword Based

12 Rule Based Extraction

12.1 Peculiarity of Numerical Relations

Analyzing a number of sentences expressing numerical relations lead to several insights as already discussed.

- **Keywords** Sentences expressing numerical relations can be expected to be explicit about the relation being expressed. Stated another way, we can expect presence of certain keywords that might help in identifying relations.
- **Modifiers** A large number of false positives stem out of mentions where a change in the numerical attribute is mentioned.

12.2 Dependencies

Dependencies are grammatical relation between two words, governer and dependent. The relation captures the way in which one of the words is affected by the other. For example, consider the sentence: "The red ball was lost" The dependencies are:

- **amod(ball,3,red,2)** "Red" is an adjective for "ball"
- **det(ball,3,The,1)** "the" is a determiner of "ball"

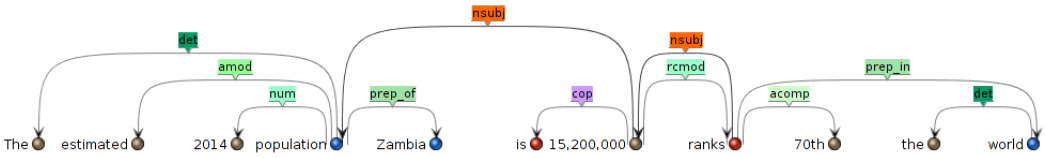


Figure 1: Dependency Graph

- **nsubjpass(lost,5,ball,3)** "ball is the subject of lost"
- **auxpass(lost,5,was,4)** "was is an auxiliary of lost"

So we know that sentences expressing numerical relations should contain express how the number and a country are related, and we know that there exists a framework which can extract pairs of words that somehow affect each other. Combining these ideas leads us to a simple rule based relation extractor.

12.3 Dependency Path

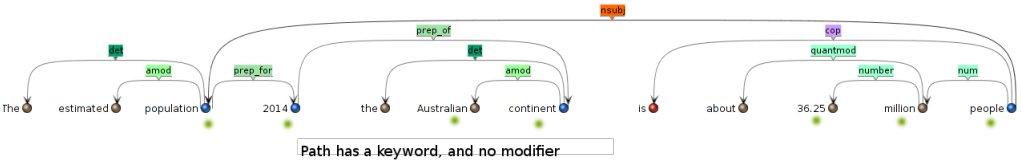
We define a dependency path between two words "A" and "B" as the shortest path between them in the dependency graph. The dependency graph consists of one node for each of the words, and the dependencies are the collapsed typed dependencies as obtained from the stanford dependency parser. [4]. The following figure shows dependency graph for "The estimated 2014 population of Zambia is 15,200,000, which ranks 70th in the world."

12.4 Relation extraction using dependency paths

Intuitively, it makes sense that the entities which are related will have some dependence on each other in the sentence expressing the relation.

12.5 Example

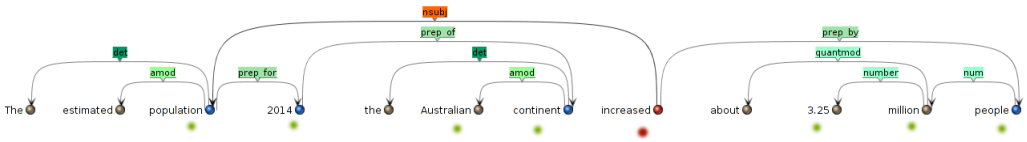
Consider the sentence: "The estimated population for 2014 of the Australian continent is about 36.25 million people" and the Country-Number pair (Australian, 36.25) The dependency graph of the sentence is as shown in figure [12.5]. The path between Australia and 36.25 has "people", so this will be an



extraction.

If we modify the sentence to: "The estimated population for 2014 of the Australian continent increased by about 3.25 million people"

The path will additionally have a modifier, "increased", and thus there won't be any extraction.



Path has a keyword: people, but also a modifier

12.6 Results

12.6.1 Precision and Recall

The extractor was applied to 30 sentences expressing 23 different relations.

	Relations Present	Relations not Present (False positives)
Extracted	16	17
Not Extracted	7	N/A

- Precision: 48.4%
- Recall: 69.6%

The precision should increase further on applying unit based pruning.

12.6.2 Example Extractions

Sentence → Extraction
At 3.71 million square miles (9.62 million km2) and with around 318 million people, the US is the world's 3rd or 4th-largest country by total area and third-largest by population. → POP(US, 318)
The land area of the contiguous US is 2,959,064 square miles (7,663,941 km2) → AGL(US, 2,959,064), AGL(US, 7,663,941)
With 1,210,193,422 residents reported in the 2011 provisional census, India is the world's second-most populous country. → POP(India, 1,028,737,436)
According to an official estimate for 1 June 2014, the population of Russia is 143,800,000. → POP(Russia, 1), POP(Russia, 2014), POP(Russia, 143,800,000)

References

[1] http://dl.acm.org/citation.cfm?id=336644Agichtein, Eugene, and Luis Gravano. "Snowball: Extracting relations from large plain-text collections." Proceedings of the fifth ACM conference on Digital libraries. ACM, 2000.

[2] Mintz, Mike, et al. "Distant supervision for relation extraction without labeled data." Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. Association for Computational Linguistics, 2009.

[3] Bunescu, Razvan C., and Raymond J. Mooney. "A shortest path dependency kernel for relation extraction." Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2005.

[4] http://nlp.stanford.edu/software/corenlp.shtml

[5] Hoffmann, Raphael, Congle Zhang, and Daniel S. Weld. Learning 5000 relational extractors. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010.

[6] http://en.wikipedia.org/wiki/Entity