Distant supervision with Units and Keywords

A) Introduction

Simple distant supervision is bound to fail on numerical attributes because of a large number of false positives.

The problem stems from the fact that numbers have no identity of their own; they represent count of some real entity or phenomenon. 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.[1]

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.

B) Role of 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.

C) Keywords

C.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.

C.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*.

[1] Some of them will be pruned by the unit extractor

* With a few exceptions, like age. For e.g. Dell, 47, lives alone.

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.

D) 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.[2]

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”

[2] A semi-successful attempt to create such keyword set using Tf-Idf scores was described in an earlier mail

E) Results

Keyword based pruning leads to a large reduction in the number of matches.

Relation	**Unit based matches**	**Unit based + keyword matches**	**Reduction%**
Land Area	98	61	37.75
GDP	1790	29	98.38
FDI	791	8	98.98
Good Export	816	20	97.549
Internet users	24369	88	99.639
Inflation	27598	981	96.445
Population	5225	961	81.608
Life Expectancy	3081	99	96.787
Co2 emission	196	6	96.939
Diesel Price	8	0	100

We manually examined all the matches for relations where only a few of the matches were left. For the others, we randomly sample matches where the total number of matches defines sample size.

We obtain some very high precision matches.

Relation	**Unit**	**Number of matches**	**Correct Matches**	**Precision (%)**
Land Area	SqKm	61	58	95.08
GDP	USD	29	1	3.4
FDI	USD	8	1	12.5
Good Export	USD	20	10	50
Internet users	%	88	20	22.72
Inflation	%	50	33	66
Population		50	42	84
Life Expectancy	Years	99	27	27.27
Co2 emission	Kiloton	20	6	30