I. Setup

Corpus	Subset of the TAC corpus, 4,499,046 sentences and 268,036 docs.
KB	Derived from worldbank, 10 attributes for different countries.
CodeBase	MultiR with matcher tweaked to identify numbers as entities. Code here.

II. Matching process

Numbers are identified by the regular expression ^[\\+-]?\\d+([,\\.]\\d+)?([eE]-?\\d+)?$

Countries are identified using a list of manually inflected country names.

A number-country pair is called a match for an attribute attr if attr(country) = number.

III. Results

a) Number of matches per relation

Attribute	Code	Matches
Land area (sq. km)	AG.LND.TOTL.K2	498
Electricity production (kWh) and Internet users %	EG.ELC.PROD.KH&&IT.NET.USER.P2	807
Pump price for diesel fuel (US$ per liter)	EP.PMP.DESL.CD	6,571
Inflation, consumer prices (annual %)	FP.CPI.TOTL.ZG	4,531
Internet users % and Land area (sq. km)	**IT.NET.USER.P2&&AG.LND.TOTL.K2**	**9**
Inflation, consumer prices (annual %) and Internet users %	IT.NET.USER.P2&&FP.CPI.TOTL.ZG	302
Internet Users %	**IT.NET.USER.P2**	**154,040**
Negative Examples	NA	39,748
Population (Total)	SP.DYN.LE00.IN	215
		206,721

b)The Match Mines

Consider the sentences in the file match_mines.txt (attached).

The number of integers and country pairs in this sentence is **very** high. South Africa will have some attribute with the value of 0, Germany will have something that is 4 and so on. With naive matching, this leaves us with a pile of false positives.

The distribution of the number of matches shows that there are 2 dangerous sentences, which are responsible for 21.8% of the matches. The top 10 of such sentences alone produce **35.7%** of the total matches, all false positives.

Sentence id	Number of matches
2300281	**22592**
2080875	**22549**
7714385	3621
6293077	3620
2655626	3617
4443497	3615
3773175	3615
3446480	3610
4464882	3594
3245301	3562

There were a total of 7,431 sentences that caused a match, and there were 206,721 matches.

The problem is that there is no dearth of sentences reporting numbers in newswire. Sports and Business sections are mostly numbers with words coated around to make them digestible.

This brings about the crucial role that units will play in this project. We cannot rely on numbers alone for matching; there are a plenty of them. We must be pragmatic and pick only those which tell us something about the attribute.

c) Some example matches (Examples with units have relevant snippets highlighted)

Country	Attribute	Value	Sentence
South Africa	Internet users %	24	Montjane, **24**, was crowned on Sunday evening at the Superbowl at Sun City, in **South Africa**'s North West province.
Lithuania	Internet users %	68	Basketball: World championships results - collated ISTANBUL, Sept 9, 2010 (AFP) World championships results on Thursday: Quarter-finals USA 89 Russia 79 Lithuania 104 Argentina 85 Played Wednesday Turkey 95 Slovenia **68** Serbia 92 Spain 89 Saturday Semi-finals USA v **Lithuania** (1600GMT), Turkey v Serbia (1830GMT) Sunday 3rd place (1600GMT) Final (1830GMT)
Nigeria	Internet users %	20	Times GMT) Jan 12 Benin v Mozambique 1830 --- Jan 16 Benin v Nigeria 1600 Egypt v Mozambique 1830 --- Jan **20** Benin v Egypt 1600 At Lubango Mozambique v **Nigeria** 1600 Note: Group winners and runners-up qualify for quarter-finals
Bangladesh	Internet Users %	7	Factbox: A/H1N1 flu cases in Asia-Pacific region HONG KONG, June 24 (Xinhua) The following are the latest confirmed cases of Influenza A/H1N1 in the Asia-Pacific region on Wednesday: Japan: 846; Australia: 2,733; South Korea: 128; China: 1044 (528 in mainland, 444 in Hong Kong, 60 in Taiwan, 12 in Macao); New Zealand: 386; the Philippines: 473; Thailand: 985; Malaysia: 80; India: 59; Singapore: 220; Vietnam: 56; French Polynesia: 1; Papua New Guinea: 1; Fiji: 3; Sri Lanka 4; Samoa: 1; Laos: 3 **Bangladesh**: **7**.
Monaco	Agricultural land area	2	Match Mine
Bermuda	Agricultural land area	50	Forecasters were also closely tracking the path of Tropical Storm Fiona about 360 miles (580 km) south of **Bermuda,** with wind speeds of up to **50** miles (85 km) per hour.
Monaco	Agricultural land area	2	Diarra 42, Gameiro 60, Monterrubio 90+4-pen) Nancy 1 (Alo'o Efoulou 90+2) Nice 1 (Remy 42) **Monaco** 3 (Nene 8-pen, Alonso 12, 71) Sochaux **2** (Brechet 3, Boudebouz 40) Valenciennes 5 (Sanchez 24, Pujol 49, 85, Audel 63, Ben Khalfallah 78) Grenoble 0 Rennes 4 (Gyan 1, 61-pen, Marveaux 7, Mangane 20) Boulogne 0 Bordeaux 2 (Gouffran 34, Lachor 70-og) Marseille 0 Montpellier 0 Sunday Lens v Lille, Toulouse v Le Mans, Paris SG v Lyon
Monaco	Agricultural land area	2	Match Mine
Benin	Electricity Production and Internet usage)	0	Match Mine. *There was a time when Benin produced no electricity and there was no one online. Benin happens to be the **only** country for which this was **ever** true. Thus, Benin is single handedly responsible for all 807 matches.*

Russia	Diesel cost	1	Svetlana Karpeeva, **Russia**, 4:44.26 Women's 4x100m freestyle relay: **1**. Svetlana Sleptsova, Russia, 43.7 (1). Anna Bogaliy-Titovets, Russia, 44.8 (1). Svetlana Sleptsova, Russia, 43.7 (1).
Afghanistan	Diesel Cost	1	Obama's strategy for **Afghanistan** and Pakistan placed the defeat of al-Qaida as the No.**1** objective, largely to make sure that the group could not plot new attacks against the United States.
Ukraine	Diesel Cost	0.3	According to latest forecast, Parts of **Ukraine** and southern Russia may receive **0.3** inch (0.8 centimeter) of rain tomorrow, and western Kazakhstan may get 0.4 inch.
France	Diesel Cost	1.72	Lotte Friis of Denmark won the women's 800 freestyle in 8:23.27, followed at 0.73 seconds by Ophelie Cyriell Etienne of **France** and Federica Pellegrini of Italy, **1.72** seconds back.
China	Diesel Cost	1.28	**China** enterprises index 1.28 pct lower -- Oct. 27 HONG KONG, Oct. 27 (Xinhua) The Hang Seng China Enterprises Index on the Hong Kong Stock Exchange fell 170.43 points, or **1.28** percent, to close Tuesday's trading at 13,145.59.
Denmark	Inflation	9	Argentina has lined up a match on Feb. 9 against England, to be played in Denmark or Norway.
Syria	Inflation	9	Syria rejects U.S. report on religious freedom DAMASCUS, Dec. **9** (Xinhua) The **Syrian** Foreign Ministry rejected the U.S. annual report released by the State Department on religious freedom, saying the report ignores that Syria is the most secular state in the region, the Syrian Damascus Press news website reported Thursday.
Syria	Inflation	3 (some other year)	1st LD: U.S. extends sanctions on **Syria** for additional year WASHINGTON, May **3** (Xinhua) Blaming Syria of posing extraordinary threat to U.S. security, President Barack Obama on Monday declared an extension of sanctions on Syria for another year.
Croatia	Inflation (Croatia indeed had an inflation of 500% in 1990, its peak being 1400% around that time)	500	**500** guns stolen from **Croatian** police depot ZAGREB, Croatia 2010-09-10 11:20:54 UTC Croatian police say more than **500** guns turned in after the country's 1991 war were stolen from a police depot and sold on the black market.
France	Life expectancy	75	**France**: Hugo Lloris, Eric Abidal, Patrice Evra, William Gallas, Bacary Sagna, Abou Diaby, Yoann Gourcuff (Florent Malouda, **75**), Jeremy Toulalan, Nicolas Anelka (Thierry Henry, 72), Sidney Govou (Andre-Pierre Gignac, 85), Franck Ribery.
France	Life expectancy	75	Bardot turns **75** with new turn in the spotlight paris, Sept 27, 2009 (AFP) Brigitte Bardot, the screen goddess who became a symbol of feminism and sexual liberation, turns **75** on Monday with her

			native **France** again honouring her with a first exhibition on her life.
France	Life expectancy	75	PLEASE SEE SEPARATE SPECIAL WORLD CUP ADVISORY FOR DETAILS Available graphics: GROUPS (8) 195 x 185 mm SQUADS (32) 45 x 185 mm STADIUMS (10) 90 x 90 mm STADIUMS MAP 175 x 140 mm KEY PLAYERS (12) 90 x 105 mm QUALIFIERS MAP 130 x 110 mm BALL 130 x 115 mm REFEREE SIGNALS 265 x 85 mm HISTORY OF RULES 265 x 110 mm AVERAGE GOALS PER MATCH 175 x **75** mm PREVIOUS WINNERS 90 x 85 mm FAKED FALLS 90 x 110 mm FBL WC2010 VUVUZELA 95 x 70 mm GOLF US OPEN graphics Text slug: Golf-USA-Open PEBBLE BEACH GOLF LINKS 130 x 85 mm PEBBLE BEACH GOLF LINKS, KEY HOLES 130 x 205 mm KEY HOLES 90 x 160 mm TOUR DE **FRANCE** Maps of stages 13-20 plus the prologue of the Tour de France cycle race in July.
Poland	Life expectancy	71	**Poland** marks **71** years since start of WWII WARSAW, Poland 2010-09-01 08:29:33 UTC Poland is marking the 71st anniversary of Nazi Germany's invasion, which started World War II.

IV) Analysis

i) Units for precision, not just recall

As numerous examples have shown, matching numbers is doomed to fail with the distance supervision assumption. MultiR relaxes this assumption (along with single relation assumption) and thus the damage done by such noisy matches may not be much.
Still, a lot of such matches can be avoided if we somehow take units into consideration. A system of matching that has sympathy with units will not only help us in getting more matches, but will also help in avoiding matches that aren't.

The subsequent points in the analysis stem from the fact that we are not catching units.

II) Dates and Sports articles lead to a lot of false positives

The top mach mines are sports articles. The table of example also contains ample sentences where we have matches due to dates. Please refer to III. b) for more on such articles.

I*II) Small numbers generate larger false positives; finer numbers generate fewer false positives.*

Intuitively, it makes sense that we'll see a lot of 2s, 3s, 71s in different contexts than 1,000,232,112.
Due to similar reasons, getting an exact match for 23.14152 is more difficult than matching 23. Note that this will not hold for distance based matching, and thus we can expect a stream of (more) false matches.

This is corroborated by unrealistically high matches for % Internet users.