I. Meta and Background

This report is a part of the series of reports describing experiments performed for numerical relation extraction using distant supervision.

We started with vanilla distant supervision, treating numbers as entities and creating a match as soon as we find a country and a number in a sentence. We observed a large number of false positives, due to the following two reasons :

a) There are simply lots of smaller numbers.

b) We were being ignorant to units.

Ignoring whole numbers smaller than hundred didn't help the situation and the number of false positives was still very high.

As the next step, we have 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::69:70;40.0::72:74;75.0:united states dollar:102:104;

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;

Which is sentence id [TAB] Number, unit, startOffset, endOffset;Number2,unit,startOff,endOff

(Code at https://github.com/NEO-IE/MultirExperiments/tree/units/MultirFramework-master/src/main/java/edu/washington/multirframework/argumentidentification/units)

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” (as defined in report 1).

II. Unit based matching with +-10% tolerance

A) Matching Technique

Two numbers are called equal if the absolute value of difference between them is less than 10% of the maximum of them.

B) Distribution of matches

AG.LND.TOTL.K2	98
BN.KLT.DINV.CD	791
BX.GSR.MRCH.CD	816
EG.ELC.PROD.KH	19
EN.ATM.CO2E.KT	196
FP.CPI.TOTL.ZG	**27598**
IT.NET.USER.P2	**24369**
NA	394
NY.GDP.MKTP.CD	1790
SP.DYN.LE00.IN	3081
SP.POP.TOTL	5225

C) Analysis

1. Units helped in fighting match mines

The number of match mines has drastically reduced. We note that the top match mine is now responsible for **only 0.15%** of the matches, as opposed to **21%** of the matches earlier.

Units indeed worked well in reducing the noise.

4852953	101
9298782	79
7179512	66
4550740	48
4031985	45
8519272	43
3004035	42

2375482	39
8217844	37
4507324	37

The new match mines are percentages. The second worst match mine is the following :

Behind Greece, Argentina's government bonds were rated highly risky by 42 percent, followed by Russia, 34 percent; Ireland, 32 percent; Portugal, 28 percent; Italy, 21 percent; Spain, 20 percent; and Mexico, 19 percent.

2. The existing methods of relaxing distant supervision assumption may break down badly.

Especially when we have smaller numbers. The extent of false positives that arise out of matching numbers may be unprecedented, and it is likely that existing "worldly entities" based techniques will break down. For example, percentage is a lost cause. Even if we find units, we will get loads of false matches for percentage, since a lot of attributes have percentage as the unit, but so do loads of other things in the world. To add to it, since it can only take a finite number of values, and with numbers beyond precision of two rare in the newswire, we will can get loads of false positives.

Is treating numbers as entities an idea that needs rethinking, especially for lower numbers? Or do we need smarter matching for them, fitting a distribution for example?

3. As a good news, there are a a fairly large number of true positives for population and land area.

4. Notes from match analysis

- When it comes to numbers, distant supervision assumption is weak as can be evident from the examples above. Since second argument of a relation is number, this number can be related to the entity in multiple number of ways, which then gives lot of false positives.

- Accuracy of the matching depends on the accuracy of the number extractor and unit extractor.
- Currently for a sentence we are using only top 1 unit given by unit extractor.
- For the attributes like i**nflation, percentage of internet user**, whose values are in percentage are affected by stock data that is heavily present in news corpus.
- Unit extractor increases the accuracy of the matches.
- Unit extractor solves the problem of small numbers (constants) which gave match mines and lot of false data.
- Total land area, population relations have lot of relevant matches.
- For relations FDI, Goods Export, GDP we have almost similar values for three values. Hence a (entity, number) pair matches all of the three relations. We can improve our

matching function for closely related attributes.

Some example matches follow. We have highlighted correct matches and matches that were wrong due to issues with Unit Tagger.

D) Examples

Country	Relation	Number	Sentence
Iceland	AG.LND.TOTL.K2	1.02999998464E11 (sq metre)	Ice and fire: recent eruptions in Iceland REYKJAVIK, April 15, 2010 (AFP) Iceland, an island of **103,000 square kilometres** (40,000 square miles) in the north Atlantic near to the Arctic Circle, has both numerous volcanoes and the largest glaciers in Europe, with ice covering around 12 percent of the country.
Lebanon	AG.LND.TOTL.K2	1.043765248E10 (sq metre)	Facts and figures on Lebanon 2009-06-07 05:49:53 UTC A look at Lebanon, which holds parliamentary elections Sunday: THE LAND -- Lebanon lies in the eastern Mediterranean and covers about **4,030 square miles** (10,450 square kilometers) -- smaller than the U.S. state of Connecticut.
Ethiopia	AG.LND.TOTL.K2	1.100000002048E12 (sq metre)	Key facts: - GEOGRAPHY: At **1,100,000 square kilometres** (440,000 square miles), Ethiopia is three times as big as Germany.
Benin, Greece	AG.LND.TOTL.K2	1.20000004096E11 (sq metre)	The United Nations this year inaugurated a decade for deserts and the fight against desertification, which affects **12 million hectares** (30 million acres) of land fit for agriculture each year, the size as Greece or Benin, Gnacadja's home country.
Mexico	AG.LND.TOTL.K2	1.999999991808E12 (sq. metre)	The wide scattering of the Marshalls' 29 atolls, 2,300 miles (3,700 kilometers) southwest of Hawaii, give them an exclusive economic zone of **800,000 square miles** (2 million square kilometers) of ocean, an area the size of Mexico.

Britain, Guniea	AG.LND.TOTL.K2	2.53818830848E11 (sq. metre)	Basic facts about Guinea CONAKRY, June 24, 2010 (AFP) Key facts about the west African state of Guinea, which holds presidential elections Sunday: - GEOGRAPHY: Bordering the north-Atlantic Ocean in tropical west Africa, Guinea covers **245,900 square kilometres** (98,000 square miles) - the same size as Britain and roughly two-thirds the size of Japan.
Guinea	AG.LND.TOTL.K2	2.53818830848E11	Guinea: part of west Africa's arc of instability conakry, Dec 4, 2009 (AFP) Key facts about the west African state of Guinea, whose military leader was shot and injured by a fellow army officer on Friday: - GEOGRAPHY: Giving on to the Atlantic Ocean in tropical west Africa, Guinea covers **245,900 square kilometres** (98,000 square miles) - the same size as Britain and roughly two-thirds the size of Japan.
Japan	AG.LND.TOTL.K2	3.99999991808E11 (sq metre)	China has never disputed Japan's sovereignty over Okinotorishima but rejects its claim to the **400,000 square kilometre** (155,000 square mile) marine footprint, which is larger than Japan's entire land area.
Lanka	AG.LND.TOTL.K2	6.4749703168E10 (sq metre)	GEOGRAPHY: Sri Lanka has a surface area of **65,000 square kilometres** (25,000 square miles).
Pakistani	AG.LND.TOTL.K2	7.99995527168E11 (sq metre)	According to Pakistani authorities, around a quarter of the country which extends over **800,000 square kilometres** (308,880 square miles) and counts 167 million inhabitants, have been affected by the floods over the last three weeks.
Brazil, Australia	AG.LND.TOTL.K2	8.514214821888E12 (square metre)	BRAZIL The following are key facts about Brazil: - GEOGRAPHY: With a total land area of **8,514,215 square**

			kilometers (3,287,357 square miles), Brazil is larger than Australia and is Latin America's biggest country.
Russia Russia China China Russia	BN.KLT.DINV.CD BX.GSR.MRCH.CD NY.GDP.MKTP.CD	1.0E10 (USD)	Russia, China seek greater international clout YEKATERINBURG, Russia 2009-06-16 12:07:30 UTC China and Russia sought greater international clout at a summit Tuesday, with China promising **$10 billion** in loans to Central Asian countries and Russia challenging the U.S. dollar's dominance as a global reserve currency.
Switzerland Switzerland Switzerland	BN.KLT.DINV.CD BX.GSR.MRCH.CD NY.GDP.MKTP.CD	1.0E10	Gates makes **$10 billion** vaccines pledge DAVOS, Switzerland 2010-01-29 18:00:26 UTC The Bill and Melinda Gates Foundation will donate **$10 billion** over the next decade to research new vaccines and bring them to the world's poorest countries, the Microsoft co-founder and his wife said Friday.
Italy Italy Italy	BN.KLT.DINV.CD NY.GDP.MKTP.CD BX.GSR.MRCH.CD	8.0E9 USD* **(We don't use precise conversion rate from euro to USD, currently it is assumed to 1)**	Italy places **8 bln euros** in bonds, rates jump ROME, Dec 30, 2010 (AFP) Italy raised more than **eight billion euros** (10.6 billion dollars) on Thursday in its final bond auction of 2010, but was forced to pay investors sharply higher rates amid persistent eurozone debt concerns.
Germany Germany Germany	BN.KLT.DINV.CD NY.GDP.MKTP.CD BX.GSR.MRCH.CD	1.0499999744E10 USD	The decision signals more EU readiness to dismantle obstacles that companies ranging from Germany's BASF to U.S.-based Monsanto have faced in seeking to expand the **$10.5 billion** global biotech crop market through growth in Europe.
Russia Russia	BN.KLT.DINV.CD NY.GDP.MKTP.CD	1.0E10 USD	Russia is spending at least **$10 billion** on dramatic upgrades to infrastructure in 13 cities to prepare to host World Cup matches, but despite the huge spending, the

Russia	BX.GSR.MRCH.CD		locals will reap the rewards for years to come, Putin said.
France, Germany, Spain France, Germany, Spain France, Germany, Spain	BN.KLT.DINV.CD NY.GDP.MKTP.CD BX.GSR.MRCH.CD	1.0E10 USD	France is the biggest beneficiary of the EU's massive farm subsidy system, the Common Agricultural Policy, receiving **10 billion euros** last year, ahead of Spain with 7.1 billion and Germany with 6.6 billion.
Ireland Ireland Ireland	BN.KLT.DINV.CD BX.GSR.MRCH.CD NY.GDP.MKTP.CD	1.0E10 USD	Under the deal, Ireland's crippled banks will immediately receive **10 billion euros** and will be subject to a fundamental downsizing, the government said.
Russia Russia Russia	BN.KLT.DINV.CD BX.GSR.MRCH.CD NY.GDP.MKTP.CD	1.0199999488E10 USD	1996: The IMF agrees to offer a loan of **US$10.2 billion** to Russia over the next three years to help Russians transform their economy.
Kuwait Kuwait, India Kuwait, India	BN.KLT.DINV.CD BX.GSR.MRCH.CD NY.GDP.MKTP.CD	1.0700000256E10 USD	In February, Bharti Airtel, India's largest telecom company, bid **$10.7 billion** for the Africa assets of Kuwait's Zain.
France	BN.KLT.DINV.CD	1.0E10 USD	France received **10 billion euros** under the CAP last year.

France France	NY.GDP.MKTP.CD BX.GSR.MRCH.CD		
Nigeria	EG.ELC.PROD.KH	1.3464000918454272E16 (joule) Unit detected: `barrel of oil equivalent`	The attacks from an insurgency that began in 2006 cut drastically into crude production in Nigeria, an OPEC-member nation that is one of the top suppliers of crude oil to the U.S. Production has risen back to **2.2 million barrels** of oil a day, in part because many militant leaders and fighters accepted a government-sponsored amnesty deal last year.
India	EN.ATM.CO2E.KT	1.259999985664E12 (kg)	Indian ports' handling capacity to reach **1.26 bln tons** by 2012 MUMBAI, India, Aug. 26 (Xinhua) - Indian ports' handling capacity will touch 1.26 billion tones by 2012 with cargo traffic at around one billion tons, said an joint report released Thursday by Federation of Indian Chambers of Commerce and Industry (FICCI) and Ernst and Young.
China	EN.ATM.CO2E.KT	5.95999981568E11 (kg)	ina's 2010 crude steel consumption to hit **596 million tonnes**: steel association BEIJING, Nov. 27 (Xinhua) China's apparent consumption of crude steel is likely to reach **596 million tonnes** this year, a year-on-year increase of 5.6 percent, according to a steel association official.
Singapore	FP.CPI.TOTL.ZG	0.6600000262260437%	Singapore shares close 0.66 percent up Singapore, Aug 14, 2009 (AFP) Singapore shares closed 0.66 percent higher Friday on gains in selected heavyweight stocks amid growing confidence about the economy, dealers said.
Singapore	FP.CPI.TOTL.ZG	0.6399999856948853 %	Singapore shares close 0.64 percent up Singapore, Sept 8, 2009 (AFP) Singapore shares closed **0.64**

	(45+ matches for Singapore and inflation)		**percent** higher Tuesday with strong buying interest in commodity-related stocks, dealers said.
Kuwait	FP.CPI.TOTL.ZG	0.6899999976158142	In Kuwait, the KSE Market or Price Index dipped **0.69 percent**, closing at 6,791.8 points, while the QE Index in Doha, Qatar, added 0.52 percent to finish at 7,593.87 points.
China	FP.CPI.TOTL.ZG	0.6999999880790771	Chinese shares were mixed Thursday -- the benchmark Shanghai Composite Index lost 20.42 points, or 0.7 percent, to 2,983.53 while the Shenzhen Composite Index for China's smaller, second exchange added **0.7 percent** to 1,248.52.
France	IT.NET.USER.P2	0.6000000238418579 (percent)	Germany's leading DAX index was down 10.27 points, or 0.2 percent, at 4,450.85 while France's CAC-40 fell 20.39 points, or **0.6 percent**, to 3,359.10.
Australia	IT.NET.USER.P2	0.6000000238418579	South Korea's main index added **0.6 percent,** China's Shanghai's benchmark climbed 1.5 percent and Australia's index advanced 1.1 percent.
Israel	IT.NET.USER.P2	20.0 %	Israel's Arab community numbers 1.3 million, about **20 percent** of the population.
China	SP.DYN.LE00.IN	1.893456E9 seconds	Such tension is routine for China's roughly eight million Uighurs, who have complained of religious oppression since the officially atheist Chinese communists came to power **60 years** ago and tightened control on Xinjiang.
Pakistan	SP.DYN.LE00.IN	1.893456E9 seconds	Why does a small elite still control vast swaths of land more than **60 years** after Pakistan became a nation?
Vanuatu	SP.DYN.LE00.IN	1.893456E9 seconds	Secretary General Kamalesh Sharma and other officials stressed that the **60-year-** old body had a

			unique role because it represented a mix of significant economic powers, such as Britain, Australia and India, as well as small and vulnerable nations such as the Maldives, Vanuatu and Tuvalu.
China	SP.POP.TOTL	9.3E8 **(Error in unit extractor)**	China and Cambodia have long had close relations, with China previously giving **930 million** dollars in aid and loans to the impoverished Southeast Asian nation since 1992, Khieu Kanharith said.
China	SP.POP.TOTL	9.5E8 **(Error in unit extractor)**	Zimbabwe receives 950 million dollar loan from China HARARE, June 30, 2009 (AFP) Zimbabwe has won **950 million** dollars in credit lines from China, the largest loan secured by the new government, boosting the country's global appeal for funds to rebuild its shattered economy.
Malaysia	SP.POP.TOTL	9163459.0 **(Error with number extractor)**	SEA Games mdals table VIENTIANE, Laos, Dec. 16 (Xinhua) Following are the medals table Wednesday at the 25th Southeast Asian Games (tabulated under country, gold, silver, bronze, total): Thailand 61 68 79 208 Vietnam 61 55 47 163 Indonesia 32 37 60 129 Singapore 32 23 33 88 Malaysia 29 36 51 116 Philippines 29 24 41 94 Laos 25 16 40 81 Myanmar **9 16 34 59** Cambodia 3 8 23 34 Brunei 1 1 8 10 East Timor 0 0 3 3
Chile	SP.POP.TOTL	8300000.0	Some **8.3 million** people are eligible to vote in Chile, one of Latin America's most prosperous nations.
Andorra	SP.POP.TOTL	82009.0	Five dead in Andorra bridge collapse ANDORRA LA VELLA, Nov **8, 2009** (AFP) Five people were killed and six injured in the collapse of a road bridge under construction in Andorra, according

III. Unit Based Exact Matching

A) Matching Technique

Two numbers are called equal if the absolute value of difference between them is less than 0.1.

B) Distribution of Number of Matches

FP.CPI.TOTL.ZG	14780
IT.NET.USER.P2	8705
NY.GDP.MKTP.CD	296
NA	268
BX.GSR.MRCH.CD	266
BN.KLT.DINV.CD	114
SP.POP.TOTL	1
AG.LND.TOTL.K2	1

We note that we don't get any matches for Population and Agricultural land area. This is however, expected. Large numbers are usually mentioned with heavy loss of digits at lower places. Eg. *"Population is 1.3 billion"*. The knowledge base on the other hand has precise facts. Since from previous analysis we know that larger numbers are rarer than smaller numbers, no exact matches for large quantities is expected.
The only exact match for Agricultural land area is correct.

C) Some Examples Matches

Country	Value	Relation	Sentence
Singapore	7e8	Agricultural land area	With a land area of just 700 square kilometres (280 square miles), Singapore does not have the watersheds and natural rivers from which to draw the life-giving resource.
Czech Republic	2.9%	Consumer price index and internet user percentage.	In negotiations to date, Britain -- together with the Netherlands, Austria, the Czech Republic, Denmark, Finland and Sweden -- have insisted the EU budget rise be limited to 2.9 percent.
Dominican Republic	1e10	GDP	After meeting with Haitian President Rene Preval and other international representatives in the neighboring Dominican Republic, Dominican President Leonel Fernandez said Haiti would need $10 billion over five years.
Ireland	1e10	GDP	Ireland is committed to slashing (EURO)10 billion ($13.2 billion) from spending and raising (EURO)5 billion ($6.6 billion) in new taxes over the coming four years, with the harshest steps coming in the 2011 budget to be unveiled Dec. 7.