sg

On Distant Supervision and its Application for Numerical Relation Extraction

October 17, 2014

Abstract

1 Introduction

2 Bare Essentials

This report will demand that the reader is on the same page vis-a-vis a few terminologies. Trading off space and brevity for clarity: a)Entity b)Entity Pair c)Relation d)Mention e)Match f)Extraction

3 Problem

The next few sections lay the foundation for discussion of our solution.

4 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.

5 Distant Supervision: Snowball scaled up

5.1 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.

5.2 Matching

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

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

5.3 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.

6 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.

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

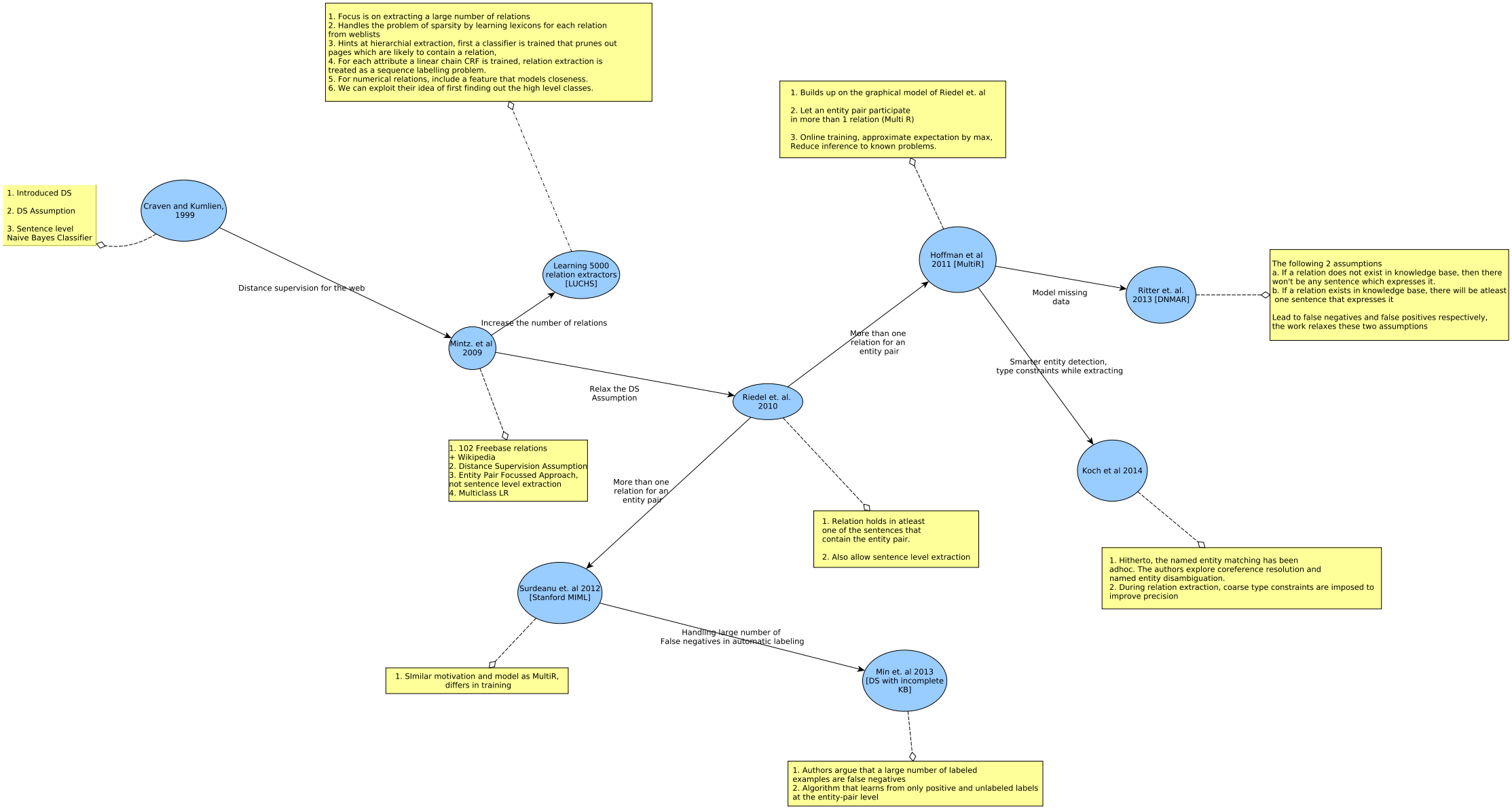
The was the first work to use distant supervision for creating a repository of biological facts. They targeted 5 different relations between Proteins, Tissues, Cell-Types, Diseases, Pharmacologic-Agents and Subcellular-structure. A naive bayes based simple relation extractor is first described. The extractor 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.

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

This work revived the interest of the community around the problem if distant supervision.


1. Focus is on extracting a large number of relations
2. Handles the problem of sparsity by learning lexicons for each relation from weblists
3. Hints at hierarchial extraction, first a classifier is trained that prunes out pages which are likely to contain a relation,
4. For each attribute a linear chain CRF is trained, relation extraction is treated as a sequence labelling problem.
5. For numerical relations, include a feature that models closeness.
6. We can exploit their idea of first finding out the high level classes.
1. Builds up on the graphical model of Riedel et. al
2. Let an entity pair participate in more than 1 relation (Multi R)
3. Online training, approximate expectation by max, Reduce inference to known problems.
1. Introduced DS
2. DS Assumption
3. Sentence level Naive Bayes Classifier
Craven and Kumlien, 1999
Learning 5000 relation extractors [LUCHS]
Hoffman et al 2011 [MultiR]
Distance supervision for the web
Increase the number of relations
Model missing data
Ritter et. al. 2013 [DNMAR]
The following 2 assumptions
a. If a relation does not exist in knowledge base, then there won't be any sentence which expresses it.
b. If a relation exists in knowledge base, there will be atleast one sentence that expresses it
Lead to false negatives and false positives respectively, the work relaxes these two assumptions
Mintz. et al 2009
More than one relation for an entity pair
Smarter entity detection, type constraints while extracting
Relax the DS Assumption
Riedel et. al. 2010
1. 102 Freebase relations + Wikipedia
2. Distance Supervision Assumption
3. Entity Pair Focussed Approach, not sentence level extraction
4. Multiclass LR
Koch et al 2014
More than one relation for an entity pair
1. Relation holds in atleast one of the sentences that contain the entity pair.
2. Also allow sentence level extraction
1. Hitherto, the named entity matching has been adhoc. The authors explore coreference resolution and named entity disambiguation.
2. During relation extraction, coarse type constraints are imposed to improve precision
Surdeanu et. al 2012 [Stanford MIML]
Handling large number of False negatives in automatic labeling
1. SImilar motivation and model as MultiR, differs in training
Min et. al 2013 [DS with incomplete KB]
1. Authors argue that a large number of labeled examples are false negatives
2. Algorithm that learns from only positive and unlabeled labels at the entity-pair level

7 MultiR

For our experiments

8 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".

9 Numerical Relation extraction using MultiR

The problem of numerical relation extraction can be easily moulded into the distant supervision framework. Recall from section [5] 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.

9.1 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.

9.2 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

10 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.

10.1 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.

11 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.

12 Numbers are weak entities: a case for keywords

12.1 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.

12.2 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.

12.3 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

13 Results

14 Possibilities

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.