A comparitive study of relation extraction algorithms using distant supervision in missing data models

Saransh Srivastava (Y9317518)
saranshs@cse.iitk.ac.in

Advisor: Dr. Arnab Bhattacharya

Department of Computer Science & Engineering,
Indian Institute of Technology, Kanpur

May 18, 2015

Outline

Introduction

- The web is filled with huge amount of free form text, which is easy for humans to read but difficult for computers to understand
- In order to share, remix and use this text information, such data needs to be organized into machine readable formats
- A novel approach of extracting relations from unstructured text corpus using *distant supervision learning algorithms*

Problem Statement

Problem Statement

Relation extraction from text corpus using distant supervision algorithm and knowledge base in missing data models

- **Distant Supervision**: is a semi-supervised learning algorithm that applies a heuristic which assumes that each sentence which mentions the two related entities is an expression of a given relation. The algorithm uses a weakly labelled training set to supervise

Problem Statement

- **Knowledge Base**:is a centralized repository to store structured information typically concepts, data, rules and specifications used by a computer system. We will use Freebase as a knowledge base. It contains approx. 44m topics and 2.4b facts.
- **Missing Data Models**: Freebase though huge, is not complete. Relations not present in database are thus treated as negative instance causing false negative in result

Motivational Example

Person	Employer
Varun Sharma	Flipkart
Naveen Tiwari	InMobi
Pawan Kumar	IIT Delhi

True Positive	"**Varun Sharma**, a manager at **Flipkart** first came up with the idea of 'customer first' in business model."
False Positive	"**Naveen Tiwari** praised **inMobi** record revenue..."
False Negative	"**Pavan Sharma**, a professor at **IIT Kanpur's** Physics Department.."

Table: A hypothetical database and heuristically labelled training data

Algorithm 1

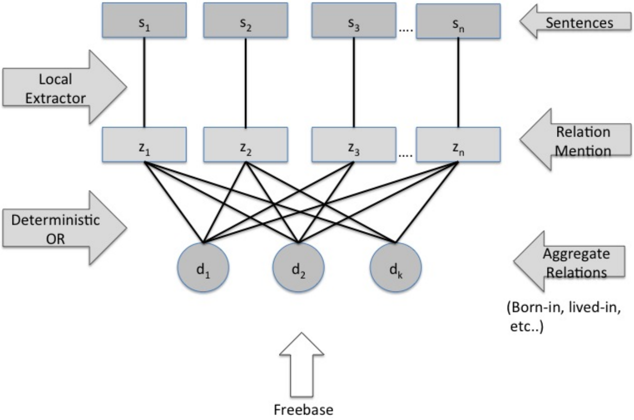


Figure: Relation extraction system

Algorithm 1

Example

"Bill Gate was the founder of Microsoft, Inc."
"Bill Gates was the CEO of Microsoft"

- Previous systems (Mintz et. al [?]) assumes that both relations do not overlap, thus cannot exists together.
- Instead distant supervision considers any one of the relationship over the entire corpus. Thus, instead of sentence level relations, the relationship is considered over the entire dataset in aggregate and is deduced by deterministic-OR
- If none of the sentences mention the relation, then the fact is considered false

Learning sentence-level relation mention

Sentence level relation mention classifier

$$\theta^* = \arg \max_{\theta} P(d|s;\theta)$$

Conditional likelihood of a given entity pair

$$P(d|s;\theta) = \frac{1}{C}\prod_{i=1}^{n} e^{\theta.f(z_i,s_i)} \times \prod_{j=1}^{k} \omega(z,d_j)$$

Learning sentence-level relation mention

Most likely sentence extraction for the label facts

$$z^{*DB} = \arg \max_{z} P(z|s, d : \theta)$$

Most likely sentence extraction, without regard to labels

$$z^{*} = \arg \max_{z} P(z, d|s; \theta)$$

Pair-wise Potential

$$\Psi(t_j, d'_j) = \begin{cases} -\alpha_{MIT} & \text{if} \quad t_j = 0 \quad \text{and} \quad d'_j = 1 \\ -\alpha_{MID} & \text{if} \quad t_j = 1 \quad \text{and} \quad d'_j = 0 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

MAP inference

- **t** (fact mentioned in text in at least one sentence) and **z** (latent sentence-level relationship mention) are deterministically related
- Maximization is now done over the additional aggregate-level variable **t**
- Ritter et al.[**?**] proposed to use greedy hill climbing method to solve using soft constraints
- Start with a random **z** and repeatedly move to the best neighbouring solution **z'**

Algorithm 2 - Passage Retrieval Model

- Extracts coarse features of the document to provide complementary feedback to information extraction model
- Passage retrieval model extracts simple lexical features like sequence of words, part-of-speech tags of words, etc..
- Xu et al.[**?**] in her model extracts Bag of words and word-position lexical features
- Combined model then predicts all relations for which the respective classifiers predicted positive results

Pseudo-relevance relation feedback

Algorithm 1 Psuedo-relevance relation algorithm

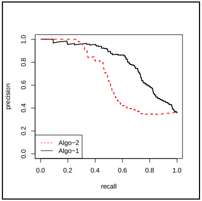
Input: Set of ground facts of relations in R
Output: Relations corresponding to the top ranked entity pairs.

initialize $\Phi' \leftarrow \Phi$
for each relationship r ϵ R **do**
    learn a passage retrieval model P(r)
        using coarse features and *PDS(r)* $\cup$ *NDS(r)*
        as training data
    score the sentences in the *RDS(r)* by P(r)
    score the pair of entities according to the scores
        of sentences they are involved in
    select the top ranked pair of entities, then add
        the relation r to their label in Φ'
end for

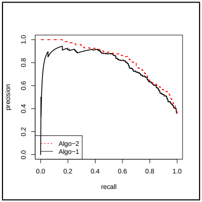
Implementation

- Detailed analysis of binary relation extraction from text corpus
- Dataset : New York Times text developed by Riedel et al. [?] containing approx. 1.8m named entities and aligned with Freebase
- Common relations extracted are compared through precision-recall values
- Implementation is in Java and Scalala programming language and precision recall graphs implemented in R
- Setup runs on Intel core i7-4770 CPU and 3.40 GHz 64-bit processor with 32 GB RAM memory in Linux environment

Overall Precision Recall Curve



(a) Without Passage-Retrieval Model



(b) With Passage Retrieval Model

Figure: Overall Precision Recall Curve at sentence-level extraction.

Overall Precision Recall Curve

- Two types of dataset are used for comparison purpose. First is the raw dataset whereas the second experiment is performed on the pre-processed dataset
- Figure(a) shows Algorithm 1 has nearly 22% more area under the curve as compared to Algorithm 2
- Figure(b) shows that Algorithm 2 performs better overall by 6.8% as compared to Algorithm 1
- Precision recall curve increases its area under curve by 26% as compared to Figure(a)

Per-relation AUC of PR curve without pre-processing

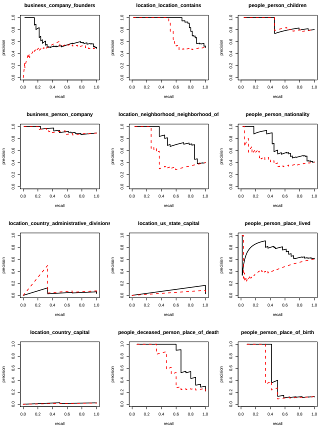


Figure: Relation extraction system

Per-relation AUC of PR curve without pre-processing

Relation	Algo 1	Algo 2
location/location/contains	0.87	0.75
people/deceased-person/ place-of-death	0.72	0.65
people/person/children	0.86	0.85
location/us state/capital	0.0	0.0
location/country/ administrative-division	0.021	0.029
location/country/capital	0.006	0.005

Table: Tabular comparison of AUC of PR curve without pre-processing

Per-relation AUC of PR curve without pre-processing

Relation	**Algo 1**	**Algo 2**
business/person/company	0.95	0.91
location/neighbourhood/ neighbourhood-of	0.55	0.50
people/person/place-lived	0.74	0.45

Table: Tabular comparison of AUC of PR curve

Per-relation AUC of PR curve with pre-processing

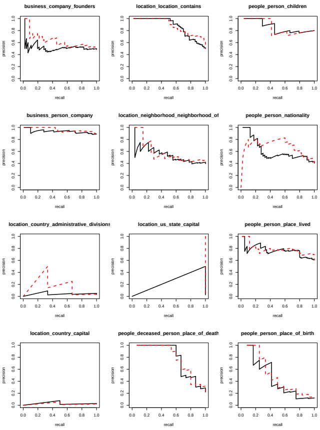


Figure: Relation extraction system

Per-relation AUC of PR curve with passage retrieval model

Relation	**Algo 1**	**Algo 2**
location/location/contains	0.85	0.86
people/deceased-person/ place-of-death	0.7194	0.7191
location/country/ administrative-division	0.02	0.06
location/us state/capital	0.0	0.0
location/country/capital	0.007	0.005

Table: Tabular comparison of AUC of PR curve with passage retrieval model

Per-relation AUC of PR curve with passage retrieval model

Relation	**Algo 1**	**Algo 2**
business/person/company	0.92	0.95
location/neighbourhood/ neighbourhood-of	0.50	0.54
people/person/place-lived	0.74	0.77

Table: Tabular comparison of AUC of PR curve with passage retrieval model

Conclusion and Future Work

- Comparative study of missing data models in relation extraction
- Algorithm 1 introduces latent variable model for learning and relaxes hard constraints
- Efficient inferencing of a large dataset is ensured via local search method and random and multiple starting points
- Algorithm 2 expands knowledge base by matching relation instance to sentences and learning the passage retrieval model and providing the relevance feedback on sentences
- These new instances are added to the knowledge base and the process is repeated to finally extract relations

Conclusion and Future Work

- Algorithm 1 is more robust and efficient in extracting relations in real time without any pre-processing
- Algorithm 2 is more accurate and performs better than Algorithm 1 but is time consuming as it required pre-processing of dataset
- Both algorithms fail to extract relations where instances are few in both text corpus and Freebase
- In future, a unification of the algorithms such that we can obtain performance level of Algorithm 2 without pre-processing of data by including techniques of Algorithm 1 will be an interesting direction to proceed

Thank You

References I

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky.
Distant supervision for relation extraction without labeled data.
In *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*, pages 1003–1011. Association for Computational Linguistics, 2009.

Sebastian Riedel, Limin Yao, and Andrew McCallum.
Modeling relations and their mentions without labeled text.
In *Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer, 2010.

A. Ritter, L. Zettlemoyer, Mausam, and O. Etzioni.
Modelling missing data in distant supervision for information extraction.
TACL, 2013.

W. Xu, R. Hoffmann, L. Zhao, and R. Grishman.
Filling knowledge base gaps for distant supervision of relation extraction.
ACL, 2013.