CS365: Discovering semantic relation in unlabeled text

Chandra Prakash pchandra@iitk.ac.in
Vishal Kumar Gupta vishalkg@iitk.ac.in
Indian Institute of Technology, Kanpur

Project Proposal
March 10, 2013

Abstract

The availability of large quantity of data due to the openess to the web text has led to the presence of a huge number of unidentified concepts and semantic relation. Most of the data that was used by advanced machine learning techniques to identify relations use human supervision for labeling them, which is not possible for huge datasets. So, there are models that can identify entities and the possible relationship between entities in an unsupervised way. In this project we plan to modify and use the Infinite Relational Model to discover semantic relationships in unlabeled text. We plan to run the algorithm on the dataset obtained by crawling the web using TextRunner and show that we can identify sensible semantic relationships between entities even with unlabeled text. We plan to compare our results with Wordnet which is a human labeled semantic lexical database for english language.

1 Introduction

One of the major goals in AI has been to automatically extract the semantic knowledge from text data, based on the relationship between concepts [3]. There has been attemps to build a question answering system by the Natural Language processing community based on the undersatnding of the given text [6]. But most of these works are not scalable because the text in these attempts have been manually tagged to identify relations and hence a large number of relations cannot be identified.

Large scale developments in Natural Language Processing has enabled huge advancement in handling a large corpus of data. Deveopments in Machine learning techniques and modern statistical approaches and the availability of huge corpus generated from automated crawling of web pages has further contributed to the building of a system that can extract semantic relation form these texts without human interference for labeling the texts for training purposes. These are mainly used by open information extractors like TextRunner which extracts large sets of triples of the form $r(a, b)$ (where r is the relation between entities a and b) in a single pass over the dataset and is purely based on unsupervised learning [4].

In some of the instances some variables or terms/codes are used to represent different objects. In these cases, even if the label for the different types of objects are available, we cannot use that label of this code words/varibles because they may refer to different objects in different situaltions. In these situations it becomes essential to identify the object using the relation in which they have been used. For example, consider the following statement "*A was expelled from the naional camp because he was caught using B for strength enhancement by C*". The relation *expelled from* and *national camp* suggest that A is an athelete. The relation *strength enhancement* suggest that B is a drug and the relation *caught by* suggest that C is police (with some probability).

We plan to approach the problem using the Infinite Relational Model [2]. This model first clusters the given entities and learns ontologies and finds the relation between the clusters that are possible. This model uses Chinese Restaurant Process (CRP) to make the number of clusters flexible. But this model uses a top down search which is not feasible for a huge dataset like the one obtained by crawling the web pages. So, there has to be a replacement of this algorithm which we have to figure out. Apart from this, we also have to find the values of suitable parameters used in CRP, Beta Distribution and the Bernoulli distribution. The dataset that we plan to use is extracted by crawling the web using TextRunner. The details of the algorithm and the dataset are specified in the remaining

sections of this project proposal.

2 Related Work

Banko *et al* [4] built a highly scalable open information extractor (OIE), that identifies relational tuples in one pass of the web pages without any supervision. It uses a self-supervised learner which produces a Naive based classifier that labels the potential extractions as reliable or non-reliable and passes this classifier to the single-pass extractor. The extractor simply uses this classifier to filter out the non-reliable extractions. Now, based on the number of occurances of the tuple in the corpus, a probability is assigned to each filtered tuple. Carlon *et al* [1] implemented NELL (never ending language learner) to read and extract information from the web to populate a growing structured knowledge base. It claims that NELL is one of the largest and most successful implementation of bootstrap learning. In 2008, Kok and Domingos [6], developed a Semantic Network Extraction (SME) model to jointly cluster the relation and object string from the tuples extracted from TextRunner [4]. The fragments of semantic networks learnt contained nodes as concept clusters which were linked with other concept clusters and the label of the links were relation clusters. In 2010, Huang and Riloff [5] built a domain-specific semantic class taggers in which the human labeled seed are used by the classifier to annotate unlabeled data which are then added to the training set and passed to a classifier for multiple semantic categories. Then the labeled texts are again passed through this classifier for cross category bootstrapping.

3 Infinite Relational Model (IRM)

Domain theory : Any set of data from a domain can be clustered into varied sets of entities. Domain theory deals in identifying those sets and discovering relations among them.

This models aims at finding effective clusters in the data set of various types so that predicting relations among the entities depends entirely on their cluster assignment. It is initially assumed that the model has access to countably infinite collection of clusters i.e. if needed the number of clusters can be increased without bounds. This assumption provides the flexibility that the number of clusters is not required to be specified in advance. We use a prior, decided through CRP, which ensures smaller number of clusters. We start with one cluster having one entity and consequently increase the number of cluster as more and more data is encountered. The two basic structures needed are the partition z and a parameter matrix η and a relation is constructed using these two. We are using conjugate priors on the entries of η. What IRM does is it inverts this relation R to find z and the parameter matrix that best fits the relation R. The above mentioned fact can be explained formally as follow: corresponding to some observed data we are having m relations $(R_1, R_2, .., R_m)$, and n types $(T_1, T_2, .., T_n)$ and let z_j be the vector of cluster assignment for the type T_j. The final goal is to find the given probability distribution: $P(z_1, ..., z_n \mid R_1, ..., R_m)$. The generative model is defined as follow [2]:

$$P(R_1, R_2, ...R_m, z_1, ...z_n) = \Pi i = 1^m P(R_i \mid z_1, ...z_n) \Pi_{j=1}{}^n P(z_j)$$

The prior probability $P(z_j)$ which we do by using CRP or Chinese Restaurant Process (CRP, Pitman 2002) also needs to be defined. In this way possible clusters are generated in the dataset. The approach which has been followed in CRP is that any cluster attracts a new object in proportion to its size. The distribution over cluster of object i is given by [2]:

$$P(z_i = a \mid z_1, ..., z_{i-1}) = \frac{n_a}{i-1+\gamma} \text{ if } n_a > 0$$
$$P(z_i = a \mid z_1, ..., z_{i-1}) = \frac{\gamma}{i-1+\gamma} \text{ if a is a new cluster}$$

where n_a is the number of objects already assigned to the cluster a and γ is a parameter. The above distribution is invariant of order and so the final prior is calculated by choosing arbitrary order and multiplying the above conditionals. To get a little more insight consider a simple case where T is 1(type: people) and a single two place relation $R : T \text{ X } T \rightarrow 0, 1$ and Ri, j is the relation showing whether i likes j or not. The complete model is represented as follow[2]:

$$z \mid \gamma \sim CRP(\gamma)$$
$$\eta(a, b) \mid \beta \sim Beta(\beta, \beta)$$
$$R(i, j) \mid z, \eta \sim Bernoulli(\eta(z_i, z_j))$$

Generalizing the above model, we get[2]:

$$R(i_1, ..i_m) \mid z_1, ..z_n, \eta \sim Bernoulli(\eta(z_{i_1}{}^{d_1}, ..z_{i_m}{}^{d_m}))$$

where d_k represents the label of the type occupying the dimension k of the m dimensional relation matrix.

Inference : To carry out the inference we used the *Markov Chain Monte Carlo* methods to sample from the posterior on cluster assignments $P(z \mid R)$ which is proportional to $P(R \mid z) * P(z)$. It can also be done by searching for the mode of distribution. During the whole process the initial clusters are modified to best fit the relation R and corresponding adjacency matrix have to be constructed to represent the relation.

4 Dataset

We are planning to use the dataset of 2.08 million tuples obtained by crawling the web using TextRunner [4], publicly available at the following url http://knight.cis.temple.edu/ yates/data/resolver_data.tar.gz. Each line in the dataset is of the form *conditions :::: , ' said :::: Dr Ghamri* where *said* is a relation over the tuple (Dr Ghamri, conditions) which are entities in this case (as interpreted by us after seeing the dataset). Apart from this dataset we are planning to use the code available at http://www.compcogscilab.com/courses/ccs-2011/ as the source code ti build the model and use this code to run the algorithm of Infinite Relational Model on the dataset.

References

[1] Carlson Andrew, Betteridge Justin, Kisiel Bryan, Settles Burr, Hruschka Jr Estevam R, and Mitchell Tom M. Toward an architecture for never-ending language learning. 2(4):3–3, 2010.

[2] Kemp Charles, Tenenbaum Joshua B, Griffiths Thomas L, Yamada Takeshi, and Ueda Naonori. Learning systems of concepts with an infinite relational model. 21(1):381, 2006.

[3] Turney Peter D, Pantel Patrick, et al. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188, 2010.

[4] Banko Michele. Open information extraction for the web. *PhD thesis, University of Washington*, 2009.

[5] Huang Ruihong and Riloff Ellen. Inducing domain specific semantic class taggers from (almost) nothing. *Proceedings of the Association for Computational Linguistics (ACL)*, 2010.

[6] Kok Stanley and Domingos Pedro. Extracting semantic networks from text via relational clustering. *Proceedings of ECML*, 2008.