International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 2 Issue: 3 627 – 630

Crime Analysis Using Self Learning

Pranav Ruke [1], Stephy Mathew[2], Meghna Mohanty[1], Pankaja Alappanavar[3,] Gandhali Gurjar[4]
Computer Department, Sinhgad Academy of Engineering [1]
IT Department, Sinhgad Academy of Engineering [2]
Assistant Professor, IT Department, Sinhgad Academy of Engineering [3]
Assistant Professor, Computer Department, Sinhgad Academy of Engineering [4]
Pune, India
rukepranav@gmail.com, stephy409@gmail.com, meghna.mohanty92@gmail.com

Abstract- An unsupervised algorithm for event extraction is proposed. Some small number of seed examples and corpus of text documents are used as inputs. Here, we are interested in finding out relationships which may be spanned over the entire length of the document. The goal is to extract relations among mention that lie across sentences. These mention relations can be binary, ternary or even quaternary relations. For this paper our algorithm concentrates on picking out a specific binary relation in a tagged data set. We are using co reference resolution to solve the problem of relation extraction. Earlier approaches co-refer identity relations while our approach co-refers independent mention pairs based on feature rules. This paper proposes an approach for coreference resolution which uses the *EM(Expectation Maximization) algorithm* as a reference to train data and co relate entities inter sentential.

Keywords-*pattern extraction, coreference resolution Introduction*

*****

I. INTRODUCTION

We consider the problem of extracting relations from huge data. Relations can be unary such as, creating just lists of various cities, movies, actors, etc. or binary such as all the (author, book) pairs.
We constructed an unsupervised machine learning process that effectively exploits statistical and structural properties of natural language discourse in order to form associations with mention pairs that do not share an identity relation. We aim to extract these unary and binary relations. We have concentrated on extracting binary relations in petty crime data obtained from newspaper articles. Our algorithm works on already tagged data that tags 'persons –locations' and 'organizations' to effectively co-relate a PERSON with its corresponding LOCATION that are connected by means of a relation type pre-defined by the user. For eg- VICTIM-POLICE STATION or VICTIM-PLACE_OF_CRIME.

For extracting inter-sentential relations, we attempt to use co reference resolution technique. Inter-sentential relations are those relations that occur across sentences. The intuition behind this is that "co-reference" can be viewed as an "identity" relation among the entity mentions."Co-reference" means co referring two or more entities and identifying a relation between them. These mentions can be present both within and across the sentences. Our approach of borrowing co-reference resolution technique for relation extraction is novel. Co reference resolution is most generally used to match and extract mention pairs with their pronoun counterpart. The seed examples are used to find pattern and features .Using these features more such examples are found out from the corpus and added to the list. These new examples are used iteratively to extract feature and add new examples

The method we describe here consists of two steps: (1)Giving the algorithm general features such as NEXT_WORD ,PREV_WORD etc to pick words that are necessary to correlate the entities. (2) self-adapting unsupervised multi-pass bootstrapping by which the system learns new rules as it reads a tagged data set using pre defined data that is hand tagged as being related or not.

When a sufficient quantity and quality of text material is supplied, the system will learn ways in which a specific set of events can be described. These mention pairs can effectively be determined as being correctly relevant based on few seed examples that will needs to be correctly hand tagged and a set of feature rules that pick out necessary words that can be correctly re iterated in the next cycle. . This method produces an accurate and highly adaptable event extraction that significantly outperforms current information extraction techniques both in terms of accuracy and robustness, as well as in deployment cost

II. BASIC CONCEPTS

Concepts of Pattern Recognition
• Pattern: A pattern is the description of an object and its occurence.
• According to the nature of the patterns to be recognized, we can divide acts of recognition into two types:
– The recognition of concrete items, for example names of people, locations etc.
– The recognition of abstract items,for example emotions, undertones of a message or article etc.

When a person perceives a pattern, he makes an inductive inference and associates this recognition with some general concepts or clues which he has derived from his past experience. The study of pattern recognition problems may be logically divided into two major categories:

– The study of the pattern recognition capability of human beings and other living organisms.
– The development of theory and techniques for the design of devices capable of performing a given recognition task for a

specific application. (Engineering, Computer, and Information Science)

Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. This activity concerns processing human language texts by means of natural language processing (NLP).

Typical subtasks of IE include:

Named entity extraction

Information Extraction. n.d. In *Wikipedia.*Wikimedia Foundation, n.d. Web.

Recognition of known entity names (for people and organizations), place names, temporal expressions, and certain types of mathematical expressions, employing existing knowledge of the domain or information extracted from other sentences in other articles. Typically the recognition task involves assigning a unique identifier to the extracted entity. A simpler task is named entity detection. Which aims to detect entities without having any existing knowledge about the entity instances. For example, in processing the sentence "Maya likes fishing", named entity detection would denote detecting that the phrase "Maya" does refer to a person, but without necessarily having (or using) any knowledge about a certain Maya who is (or, "might be") the specific person whom that sentence is talking about.

Relationship extraction

Identification of relations between entities, such as: PERSON works for ORGANIZATION (extracted from the sentence "John works for Tata.")PERSON located in LOCATION (extracted from the sentence "Nancy is in Belgium.")

Bootstrapping

Bootstrapping can start either with a set of predefined rules or patterns, or with a collection of training examples (seeds) annotated by a domain expert on a (small) data set. These are normally related to a target application domain (in our case on crime corpus)and may be regarded as initial “rules” to the learning system. The training set enables the system to derive initial extraction rules, which are applied to tagged data set in order to produce a much larger set of examples. When the new rules are subsequently applied to the text corpus, additional instances of the target will be identified, some of which will be positive and some not. As this process continues to iterate over, the system acquires more extraction rules, fanning out from the seed set until no new rules can be learned. Thus defined, bootstrapping has been used in natural language processing research, notably in word sense disambiguation (Yarowsky, 1995). Strzalkowski and Wang (1996) were first to demonstrate that the technique could be applied to adaptive learning of named entity extraction rules. In this paper, we describe a different approach on building event patterns and adapting to the different structures of unseen events. For eg we use a feature called NEXT_WORD that picks the next word that immediately follows the word tagged as /PERSON using the tagged seed examples. Our algorithm learns if a person is followed by LRB-AGE-LRB then it learns that a person is most probably a victim.

III. NEED

The proposed work, an unsupervised algorithm, extracts the required information from the corpus. Some small number. of seed examples are used. These seed examples are used to find pattern and features .Using these features more such examples are found out from the corpus and add to the list. These new examples are used iteratively to extract feature and add new examples. Here, the relationships are found out which may be spanned over the entire length of the document.

This project will enable organizations looking to parse large amounts of data with less effort and less time. Currently it can be used by police departments to keep track of criminals, areas where crime occurs most frequently and modus operandi of criminals. The scope can further be increased to keep track of police stations and inspectors.

We are also using agriculture information to check for most frequently occurring diseases and cures in crops like rice, soyabean and cotton, such information will be valuable to farmers and the agriculture industry to keep a check of disease and cures of various crops.

IV. MATHEMATICAL MODEL

Given a document DOC consisting of n relations, $r_1, \ldots, r_n$, we use Pairs(DOC) to denote the set of relation pairs, $\{r_{ij} \mid 1 \le i < j \le n\}$, where r_{ij} is formed from relations r_i and r_j.

The pairwise probability formed from r_i and r_j refers to the probability that the pair r_{ij} satisfies the given relation(VICTIM-POLICE_STATION and VICTIM-PLACE_OF_CRIME) and is denoted as Pcoref (r_{ij}).

A clustering of n relations is an (n x n) Boolean matrix C, where C_{ij} (the (i,j)th entry of C) is 1 if and only if relations r_i and r_j satisfies the given relation. An entry in C is relevant if it corresponds to a relation pair in Pairs(DOC).

The Model

The generative model operates at the document level, inducing a valid clustering on a given document DOC. More specifically, our model consists of two steps. It first chooses a clustering C based on some clustering distribution P(C), and then generates DOC given C:

$$P(DOC,C) = P(C)P(DOC \mid C).$$

To facilitate the incorporation of linguistic constraints defined on a pair of relations, we represent D by its relation pairs, Pairs(DOC). Now, assuming that these relation pairs are generated conditionally independently of each other given C_{ij},

$$P(DOC \mid C) = \prod_{r_{ij} \in \text{ Pairs(DOC)}} P(r_{ij} \mid C_{ij}).$$

Next, we represent r_{ij} as a set of features that is potentially useful for determining whether r_i and r_j are coreferent. Hence, we can rewrite P(DOC | C) as

$$\prod_{r_{ij} \in \text{ Pairs(DOC)}} P(r1_{ij}, \ldots, r11_{ij} \mid C_{ij}),$$

where rk_{ij} is the value of the kth feature of r_{ij} .

We use 11 most generic features for our data set as shown in table 1.1. Hence we consider relation pairs upto r11.To reduce data sparseness and improve the estimation of the above probabilities, we make conditional independent assumptions about the generation of these feature values.
We have formed a set of generalised features as shown in Table 1.1

Feature ID	Feature Name	Feature Description
1	Acrsent	Do entities lie across sentences
2	nextp	Word after person
3	nextl	Word after location
4	prevl	Word before location
5	prevp	Word before person
6	Vb4l	Verb before location
7	Vb4p	Verb before person
8	Vafl	Verb after location
9	Vafp	Verb after person
10	bwloc	Location between person and location
11	bwper	Person between person and location

Table 1.1

These features pick up words and check whether the current relation pair has the same values as that of the tagged data set.

The Induction Algorithm

To induce a clustering C on a document DOC, we run EM on our model, treating DOC as observed data and C as hidden data. Specifically, we use EM to iteratively estimate the model parameters, Θ, from documents that are probabilistically labeled (with clusterings) and apply the resulting model to probabilistically re-label a document (with clusterings). More formally, we employ the following EM algorithm:

E-step: Compute the posterior probabilities of the clusterings, P(C|D, Θ), based on the current Θ.

M-step: Using P(C|D, Θ) computed in the E-step, find the Θ ′ that maximizes the expected complete log likelihood, ∑C P(C|D, Θ) log P(D,C| Θ′).

The induction process starts at the M-step. To find the Θ that maximizes the expected complete log likelihood, we use maximum likelihood estimation with add-one smoothing. Since P(C|DOC, Θ) is not available in the first EM iteration, we instead use an initial distribution over clusterings, P(C). Now which P(C) to use? One possibility is the uniform distribution over all (possibly invalid) clusterings. Another choice is a distribution that assigns non-zero probability mass to only the valid clusterings. Yet another possibility is to set P(C) based on a document labeled with coreference information. In our experiments, we use this last method .Another possibility is to begin at the E-step by making an initial guess at Θ. a probability of one to the correct clustering of the labeled. After (re-)estimating Θ in the M-step, we proceed to the E-step, where the goal is to find the conditional clustering probabilities. Given a document DOC, the number of coreference clusterings is exponential in the number of mentions in DOC, even if we limit our attention to those that are valid. To cope with this computational complexity, we approximate the E-step by computing only the conditional probabilities that correspond to the N most probable coreference clusterings given the current Θ. We identify the N most probable clusterings and compute their probabilities. To obtain the required conditional clustering probabilities for the E-step, we normalize the probabilities assigned to the N-best clusterings so that they sum to one.

For our project we consider a set of 10 seed examples- this depends on the size of the actual data set and seed examples would vary accordingly. We manually tag related relations as (1,0) and non related relations as (0,1). The algorithm then considers those relations which are tagged as (1,0) and picks out features from the feature list and when it comes across an unlabelled paragraph will calculate a probability by which this unlabelled paragraph would be related or not based on the features it has extracted and similarity to a seed example. Thus an unlabelled paragraph now has a weighted probability of being related or not.

V. RESULT

When our algorithm runs on an unlabelled data set it shows a precision of 84% and a recall of 94%.
Thus 84% of the untagged paragraphs on a crime corpus are correctly tagged by our system.
Since our project uses generic algorithm which can be applied over to a wide variety of data by just changing the seed tuples, it has a lot of application in different fields as it allows computer to understand and use large amount of data. This it can be used on data sets relating to politics, sports or general crime to be used to predict unary and binary relations among entities occurring in such articles.

VI. CONCLUSION

Nowadays internet is becoming the main source of information. But the information is scattered, it would be useful if it is in a concise and more usable form. Most of the data on internet is given in the form of textual data and we have to manually read to get the information we want but with this system reduces the human effort as it extracts the useful information in form of relation tuples which can be directly fed into further applications. Since the data is in tabular format complex queries can be applied to it to get more precise data. For example a corpus of chain snatching crimes has been used to test the system. The system extracts information from the corpus like victim and police station; this can be used in an application to check which area has maximum occurrence of the crime.

Our goal is to expand our extraction algorithm to extract ternary as well as quaternary pairs. We also aim to compute more features to improve the precision and recall of the algorithm.

This system can also be used to create large databases by using only a few seed examples. This system can also stand as reference for predictive crime analysis by analyzing and extracting most frequently occurring petty crime sites and timings.

ACKNOWLEDGMENT

The authors would like to thank TRDDC for giving us the opportunity to work on this project. We would like to show our sincere gratitude to our guide Prof. G. S. Gurjar and Prof P. B. Alappanavar for their guidance and knowledge without which this paper would not be possible. They provided us with valuable advice which helped us to accomplish writing this paper. We are also thankful to our HOD Prof. B. B. Gite (Department of Computer Engineering) and HOD Prof. Abhay Adapanawar (Department of IT Engineering) for their constant encouragement and moral support.

Also we would like to appreciate the support and encouragement of our colleagues who helped us in correcting our mistakes and proceed to complete the paper with the required standards.

REFERENCES

[1] S. Brin. Extracting Patterns and Relations from the World Wide Web. In WebDB Workshop at 6th International Conference on Extending Database Technology,EDBT'98, pages 172–183, Valencia, Spain, 1998.

[2] Banko, Michele, et al. "Open Information Extraction from the Web." IJCAI.Vol. 7. 2007.

[3] Chu, Eric, et al. "A relational approach to incrementally extracting and querying structure in unstructured data." Proceedings of the 33rd international conference on Very large data bases.VLDB Endowment, 2007.

[4] Dalvi, Bhavana Bharat, William W. Cohen, and Jamie Callan. "Websets: Extracting sets of entities from the web using unsupervised information extraction." Proceedings of the fifth ACM international conference on Web search and data mining.ACM, 2012.

[5] Mooney, Raymond J., and RazvanBunescu. "Mining knowledge from text using information extraction." ACM SIGKDD explorations newsletter 7.1 (2005): 3-10.

[6] Etzioni, Oren, et al. "Open information extraction: The second generation." Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume One.AAAI Press, 2011.

[7] Culotta, Aron, Andrew McCallum, and Jonathan Betz. "Integrating probabilistic extraction models and data mining to discover relations and patterns in text."Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics.Association for Computational Linguistics, 2006.

[8] Bollegala, DanushkaTarupathi, Yutaka Matsuo, and Mitsuru Ishizuka. "Relational duality: Unsupervised extraction of semantic relations between entities on the web." Proceedings of the 19th international conference on World wide web. ACM, 2010.

[9] Akbik, Alan, et al. "Unsupervised Discovery of Relations and Discriminative Extraction Patterns." COLING. 2012.

[10] Yan, Yulan, et al. "Unsupervised relation extraction by mining Wikipedia texts using information from the web." 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.

[11] Luo, Xiaoqiang, et al. "A mention-synchronous coreference resolution algorithm based on the bell tree." Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004.

[12] Mohanty, Meghna, Pranav Ruke, Stephy Mathew, Gandhali Kulkarni, and Pankaja Alappanavar. "Unsupervised Relation Extraction.

[13] Ng, Vincent. "Unsupervised models for coreference resolution." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008.

[14] Yarowsky, David. "Unsupervised word sense disambiguation rivaling supervised methods." Proceedings of the 33rd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1995.

[15] Strzalkowski, Tomek, et al. "Natural Language Information Retrieval: TREC-5 Report." TREC. 1996.

[16] Strzalkowski, Tomek, Jin Wang, and Bowden Wise. "A robust practical text summarization." Proceedings of the AAAI Symposium on Intelligent Text Summarization. 1998