ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

Unsupervised Relation Extraction

Meghna Mohanty, Pranav Ruke, Stephy Mathew ,Gandhali Kulkarni[1], Pankaja Alappanavar[2]

Professor, Computer Department, Sinhgad Academy of Engineering, Pune, India [1]

Assistant Professor, IT Department, Sinhgad Academy of Engineering, Pune, India [2]

Abstract: - World Wide Web consists of vast information which is scattered across millions of web pages. We consider the problem of extracting relations from this 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 want to propose an unsupervised algorithm to extract the required information from the corpus. Some small no. of seed examples can be used. Here, we are interested in finding out relationships which may be spanned over the entire length of the document. It is important to note that, our algorithm differs from the previous algorithms proposed in following aspects.

1. Rather than a sentence level or limited context, we would like to consider larger context, i.e. whole document.
2. Also, we would like to extend to higher order relations, i.e. extracting n-tuples, where, n > 2.

Keywords: Relation extraction, event extraction, pattern extraction, unsupervised

I. INTRODUCTION

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).Applying information extraction on text, is linked to the problem of text simplification in order to create a structured view of the information present in free text.

Typical subtasks of IE include:

1. Named entity extraction

Recognition of known entity names (for people and organizations), place names, temporal expressions, and certain types of numerical expressions, employing existing knowledge of the domain or information extracted from other sentences. 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 "M. Smith likes fishing", named entity detection would denote detecting that the phrase "M. Smith" does refer to a person, but without necessarily having (or using) any knowledge about a certain M. Smith who is (or, "might be") the specific person whom that sentence is talking about.

2. Relationship extraction: identification of relations between entities, such as: PERSON
3. works for ORGANIZATION (extracted from the sentence "Bill works for IBM.")PERSON located in LOCATION (extracted from the sentence "Bill is in France.")

II. NEED

The proposed work, an unsupervised algorithm, extracts the required information from the corpus. Some small no. 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.

III. BASIC CONCEPTS

Concepts of Pattern Recognition

• Pattern: A pattern is the description of an object.

• According to the nature of the patterns to be recognized, we may divide our acts of recognition into two major types:

– The recognition of concrete items

– The recognition of abstract items

When a person perceives a pattern, he makes an inductive inference and associates this perception with some general

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

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. (Psychology, Physiology, and Biology)
– 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)

IV. APPLICATION

Unsupervised Relation Extraction approach results in fast and high-precision named entity recognition, since one simply looks for occurrences of any entries in the gazette - though occasionally one needs some post-processing to distinguish between an occurrence of London as CITY or as PERSON (e.g., Jack London). But the accuracy (recall) of this approach is critically dependent on the completeness of the algorithm used.

Also this system can be used to create large databases by using only a few seed examples. For example many books like “Disbanded" by Douglas Clark which was published online is not present in any online sources. If the book list can be expanded and if almost all books listed in online sources can be extracted, the resulting list may be more complete than any existing book database. The generated list would be the product of thousands of small online sources as opposed to current book databases which are the products of a few large information sources. Such a change in information flow can have important social ramifications.

4.1 Generating Useful databases for Institutes

The algorithm can be applied over to a wide variety of data by just changing the seed tuples. Thus the system reduces human effort as it extracts the useful information from textual data 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

4.2 Creating Patterns

The information scattered over the internet can be extracted under one domain and patterns can be formed from the information extracted. For example, a database can be created for chain snatching using news articles from the internet. This can be provided to the police department and they can find patterns, as to where and when the crime takes place and they can take required actions.

4.3 System For Resume Processing

Repositories of resumes, of both external candidates and current employees, contain valuable information. The information and insights mined from resume repositories can be used to improve the quality of employees’ work, satisfaction levels for employees, as well as improvements in HR processes. Specifically, these insights can help in various practical tasks such as forming the right team for a project, selection of right candidates for a particular job requirement helping in career path planning, identifying right training programmes and reducing recruitment costs.

4.4 Solutions for Agriculture Industry

A corpus of Agricultural information can be used. The system extracts information from the corpus like crop, disease and treatment; this information can be used in an application to check which crop is subjected to maximum occurrence of a particular disease and proper treatment or precautionary measures can be taken.

V. DOCUMENT PRE-PROCESSING

There is a need to perform certain cleanup and other pre-processing operations on the documents in the given corpus, before they can be given as input to the unsupervised relation extractor algorithm. The pre-processing includes spelling corrections, format conversion (e.g., Microsoft Word to plain text), adding markers to preserve information (e.g., bold, italics, underline, font size), sentence boundary detection, expansion of abbreviations such as Co. to Company, Ltd. to Limited, removing unwanted symbols/words/characters (e.g., replacing I.B.M. with IBM or replacing !!! with !) etc. Each of these pre-processing steps is implemented as a regular expression.

VI. PROJECT PLAN

In the proposed system we aim to build a software that can extract entities from data supplied through features that we give through some seed examples. This system is semi-supervised or unsupervised which implies that that it has to itself recognized named entities based only on seed tuples and features provided in the data..

The following figure (fig. 2) depicts the project plan. It describes the activity plan of the project. The activities will be carried out in the same order.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

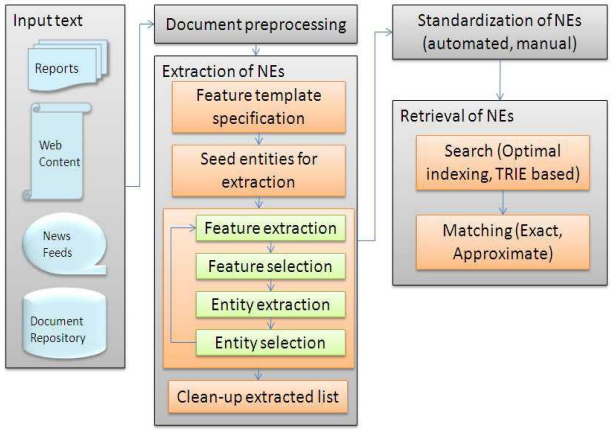

Fig. 2Project Plan

VII. CONCLUSION

The project uses generic algorithm which can be applied over to a wide variety of data by just changing the seed tuples. Thus the system has a lot of application in different fields as it allows computer to understand and use large amount of data.

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.

Also this system can be used to create large databases by using only a few seed examples. For example many books like "Disbanded" by Douglas Clark which was published online is not present in any online sources. If the book list can be expanded and if almost all books listed in online sources can be extracted, the resulting list may be more complete than any existing book database. The generated list would be the product of thousands of small online sources as opposed to current book databases which are the products of a few large information sources. Such a change in information flow can have important social ramifications. Much scope for future work remains – virtually every aspect of our system can be "drilled down" upon to discover and evaluate alternative approaches. Some outstanding problems include:

• How to handle updates to the unstructured data. That is, how should these updates be propagated to the wide table, the mapping table, etc.?

• How to record the evolution of data, so when a new document arrives and we find that it is similar to existing documents in the workbench, we know how we should process the new document.

VIII. ACKNOWLEDGEMENT

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.

IX. 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.