Information Extraction from Unstructured Text Data-A Review

Jancy Joseph, *Asst.Professor, Dept. of Computer Application, St.Joseph's College Pilathara, Kannur, Kerala, India*

Abstract

A huge volume of unstructured data is being produced in digital form in the current era. It's a laborious task to get relevant information from this unstructured data and it take much time and effort to read all these texts manually. Information Extraction (IE) system help to extracts useful information from this huge volume and variety of unstructured data. This study reviews various existing IE techniques, limitations and challenges in Information Extraction from unstructured text-based data. Information Extraction systems takes natural language text as input and produces structured information. Named entity Recognition (NER), Relation Extraction (RE), Event Extraction (EE) are the various subtasks of IE.

Keywords: Information Extraction, Named entity, NER, Relation Extraction, Event extraction

Introduction

Information Extraction (IE) is the process which extracts structured information from unstructured data.IE extracts structured content in the form of entities, relations, facts and other type of information. It is the task of automatically extracting structured information from unstructured or semi-structured data.IE technology is concerned with structuring relevant information from a text of given domain. The goal of an IE system is to find and link the relevant information while ignoring the extraneous and irrelevant data (Tellez-Valero, Montes-y-Gomez, & Villasenor-Pineda, 2005). This extracted data helps to prepare data in data analysis. The efficient and accurate transformation of unstructured data leads to improved performance of data analysis and IE.IE systems are based on NLP, language modelling, and structure extraction techniques. A huge amount of unstructured data are available in digital form on the internet and intranets like government documents, corporate reports, online news, court rulings, legal acts, medical alerts and records, and social media communication. This led to the need of effective and efficient techniques for analysing free text data and discovering valuable and relevant knowledge from it in structured manner and led to the emergence of Information Extraction technologies (Varsha Pande, 2016).

Challenges in IE

The extraction of structured data from noisy, unstructured sources is a challenging task.IE is easier for languages like English and Russian but it is difficult for agglutinative language like Malayalam because of its undefined structure. Big data adds more challenges to the IE process due to huge volume and variety of data. Data can be structured, semi-structured and unstructured. The unstructured data have no schema and have multiple format. This data come from diverse sources and there is no standardization for data. So unstructured data can have different representations. The complicated heterogeneity of mixed data make it difficult to extract useful information. Huge volume of multifaceted unstructured data make IE process more challenging (Kiran Adnan, Rehan Akbar, 2019). Scalability, dimensionality, and heterogeneity of unstructured data appear as main challenges to extract useful information from unstructured data. Unstructured data can be in the form of text, image, audio, and video format.

Applications of IE

Textual sources from which information extraction can retrieve structured information are legal acts, medical records, social media interactions and streams, online news, government documents, corporate reports and more. Information Extraction is useful in various scientific, personal, enterprise applications. Information extraction can be applied to a wide range of textual sources from emails and Web pages, reports, presentations, legal documents and scientific papers. The technology successfully solves challenges related to content management and knowledge discovery in the areas of:

- Business intelligence (for enabling analysts to gather structured information from multiple sources);
- Financial investigation (for analysis and discovery of hidden relationships);
- Scientific research (for automated references discovery or relevant papers suggestion);
- Media monitoring (for mentions of companies, brands, people);
- Healthcare records management (for structuring and summarizing patients' records);
- Pharma research (for drug discovery, adverse effects discovery and clinical trials automated analysis).
- News Tracking-automatically tracking specific event types from news sources.

Literature Review

A preliminary study has been conducted to know the existing literature on the topic Information Extraction from unstructured data. It helped to know the currently existing techniques and methods used for IE. The review helps to know the limitations of existing IE system and challenges faced in this area. This study included reviews on various IE techniques to extract relevant data from unstructured text data. The journals referred in the present study are International Journal of Engineering Business Management, Journal of Big Data, IOSR-JCE, ARPN Journal of Engineering and Applied Science, AI Magazine, IJCSI International Journal of Computer Science Issues and International Journal of Scientific & Engineering Research.

Methodology

IE procedures emerged in 1990s. In early stages, IE systems worked based on template filling, rule-based methods, classification model-based methods and sequential labelling.IE systems are based on NLP, language modelling, and structure extraction techniques.

Information Extraction from Text

IE process identifies and represents structured information from natural language text. Text strings, values or tags extracted from the text are specified in user-defined structure called templates or objects. (Kiran Adnan, Rehan Akbar, 2019). Information extraction task has been divided into two sub tasks like the high level structure and the low level structure. High level structure includes major parts of the text documents such as heading and title. Low level structure includes named entities, events, relation and terms which help to understand the content and context of the text document and it is more important and complex task. Segmentation, Classification, association, normalization and de-duplication are five tasks for general IE and storage. Machine translation, auto-coding, indexing and term extraction are the main techniques to give meaning to unstructured data in IE process. Auto-coding and indexing help to identify terms from text (Kiran Adnan, Rehan Akbar, 2019). Information Extraction from Text can be mainly divided into three subtasks such as Named Entity Recognition, Relation Extraction and Event Extraction.

Named Entity Recognition (NER)

NER is used to extract descriptive information from entities such as person names, location, organizations, numbers, currency etc. Entities can be generic such as person and location or domain specific like proteins, chemicals, drugs etc. The NE tasks was first introduced as part of MUC 6(Message Understanding Conference) evaluation exercise in 1995 and was continued in MUC 7 in 1998. They formulated 7 types of NE: PERSON, ORGANIZATION, LOCATION, DATE, TIME, MONEY and PERCENTAGE (Bindu M.s, Sumam Mary Idicula, 2011).Identification of entities (Named Entity detection) and their classification (Semantic classification) are the subtasks of NER. Modern solutions to NER are based on statistical sequence labelling algorithm. Extraction models in NER system uses three techniques

- Rule-based methods
- Machine Learning Algorithm
- Hybrid approaches

Rule-based methods for NER use Lexico-syntactic patterns and semantic constraints to identify the occurrence of similar entities. Learning based methods use machine learning to extract named entities and their classification. Learning based methods can be supervised, unsupervised (clustering), semi supervised (bootstrapping) learning. Hybrid approaches achieve better performance and accuracy. (Kiran Adnan, Rehan Akbar, 2019).

Machine learning approach is best suited for NER techniques for various Indian regional language like Malayalam, Hindi, Telugu, etc. (Kiran Adnan,Rehan Akbar, 2019).

The Named Entity Recognition (NER) term is proposed and construed in the Message Understanding Conference (MUC-6) that discussed the IE area and put forward the IE research methods. (Muawia Abdelmagid, 2015).

NER are entity specific and technique-related. Traditional NER techniques are inadequate to handle the dimensionality and heterogeneity of unstructured big data. Supervised learning techniques require large annotated data for training and that is a laborious and difficult task for large scale data sets. Weakly supervised learning is effective as compared to supervised due to reduced manual effort. But still, shortage of data make these techniques incompetent.

Major entity-specific challenges of NER are open nature of vocabulary, abbreviations, disambiguation, different languages and domain (Kiran Adnan, Rehan Akbar, 2019). Identification of named entities from an open domain unstructured text is a challenging task. Identification and classification of named entities to semantically meaningful classes is the ultimate goal of named entity recognition systems. The named entity recognition in

Malayalam is a challenging task due to the following set of reasons. Languages like English has a capitalization feature for identifying named entities. But Malayalam doesn’t have capitalization feature which makes it more challenging to identify named entities. Rich morphology of Malayalam is another problem in identifying named entities (Ajees A Pa, Sumam Mary Idiculaa, 2018) (Ajees A Pa, Sumam Mary Idiculaa, 2018)

Relation Extraction

Relation Extraction (RE) is the task of extracting semantic relationships from text, which usually occur between two or more entities. These relations can be of different types.

E.g “Paris is in France” states a “is in” relationship from Paris to France. It is the subtask of IE. Relation Extraction is the process of finding the semantic relation between entities from text. The system necessitates to correctly annotate the data by identifying a piece of text having the semantic property .Different techniques used to extract relation between identified entities are knowledge-based methods and supervised methods (Kiran Adnan, Rehan Akbar, 2019). The relation extraction task can be divided into two steps.

- Detecting if a relation word corresponding to some entity mention pair of interest occurs
- Classifying the detected relation mentions into some predefined classes.

There are two types of relation Extraction systems:-Closed domain relation extraction system and open domain relation extraction system. (Sing, 2018). Closed domain relation extraction systems consider only a closed set of relationship between two arguments. While open-domain relation extraction systems use an arbitrary phrase to specify relationship. We can extract relations by analysing the sentences in the free text using POS tagger, dependency parser and a NER. Early RE approaches can be categorized into feature based methods and kernel-based methods. (Sing, 2018). Other approaches used for RE are Lexico-Syntactic pattern or Hearst Patterns as syntactic features, and use semantic features. In 2016 Tesfaye et al., proposed a hybrid approach to get semantic information extracted from the Wikipedia hyperlink hierarchy of the constituent words. Recently, relation extraction models based on deep learning have achieved better performance than conventional relation extraction models that rely on hand-crafted features (Sing, 2018).

Event Extraction

Event Extraction is the process of gathering knowledge about periodical incidents found in texts, automatically identifying information about what happened and when it happened (Nader, 2019). An event consists a trigger or arguments. A trigger is a verb or normalized verb that denotes the presence of an event in the text. Arguments are the entities which assign semantic roles to illustrate their influence towards event description. In Event Extraction, event detection is performed. This step is frequently divided into two separate stages: trigger detection, which consists of the identification of event triggers and their type, and edge detection (or event construction), which is focused on associating event triggers with their arguments (Jorge A. Vanegas, Sérgio Matos, Fabio González, and José L. Oliveira, 2015). Different kinds of techniques for event extraction are there like data-driven, knowledge-driven and hybrid approaches. Data –driven approaches uses word, n-grams and weight while Knowledge-driven approaches use lexico-syntactic patterns and lexico-semantic patterns. Data-driven approaches need more data as input with less domain knowledge whereas knowledge-based methods require little data but high knowledge and expertise (Kiran Adnan, Rehan Akbar, 2019).Hybrid approaches compromise the effort and improve the performance.

Conclusion and Future Scope for Research

Information Extraction is a promising field in NLP due to the rise of social media platforms such as twitter, Instagram, online news channels etc. IE systems try to extract the semantics of text in natural language relationships, but most systems use supervised specific examples of the relationship to learn. The extracted information can be used for many downstream tasks including text summarization, event extraction, conceptual mapping and relation extraction (Thornton, 2019).IE system has progressed with variety of IE techniques such as open Information Extraction (OpenIE), semi-structured extraction etc. Open IE systems such as TextRunner, show unlimited number of relationships found on the Web to handle (Talha Mahboob Alam,Mazhar Javed Awan, JUNE 2018) . A rapid hike in data has occurred due to the plenty use of social media platform in this time of Covid-19 pandemic. Social media reacts to world events faster than traditional news sources. Because of the existence of abundant data, IE comes in big relevance. This study is focussed to reach to various approaches and methods to do extraction of relevant data. This review was conducted to investigate the effectiveness and limitations in existing IE techniques and the researcher tried to find out new opportunities and challenges in IE system.

References

Ajees A Pa, Sumam Mary Idiculaa. (2018). A Named Entity Recognition System For nMalayalam Using Neural Network. *8th International Conference on Advances in Computing and Communication (ICACC-2018).* Elsevier.

Bindu M.s, Sumam Mary Idicula. (2011). Named EntityIdentifier for Malayalam Using Linguistic Principles Employing Statistical Methods. *IJCSI Indernational Journal of Computer Science Issues*, 185-190.

Jorge A. Vanegas, Sérgio Matos, Fabio González, and José L. Oliveira. (2015). An Overview of Biomolecular Event Extraction from Scientific Documents. *Hindawi Publishing Corporation Computational and Mathematical Methods in Medicine*, 19.

Kiran Adnan, Rehan Akbar. (2019). Limitation of Information Extraction methods and techniques for heterogeneous unstructiured big data. *Internatioinal Journal of Engineering Business Management*, 1-23.

Kiran Adnan,Rehan Akbar. (2019). An analytical Study of information extraction from unstructured and multidimensional big data. *Journal of Big Data*, 6-91.

Muawia Abdelmagid, A. A. (2015). Information Extraction Methods and Extraction Techniques in the Chemical Document's Contents:survey. *ARPN Journal of Engineering and Applied Sciences*, 1068-1073.

Nader, R. (2019, May 3). Natural Language Processing — Event Extraction. *Towards Data Science*.

Sing, S. (2018). *Natural language Processing for information Extraction.* Cornel University: arXiv cs.CL lab.

Talha Mahboob Alam,Mazhar Javed Awan. (JUNE 2018). Domain Analysis of Information Extraction Techniques. *INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY SCIENCES AND ENGINEERING, VOL. 9*.

Tellez-Valero, A., Montes-y-Gomez, M., & Villasenor-Pineda, L. (2005). A Machine Learning Approach to Information Extraction. *Proceedings of the 6th international conference on Computational Linguistics and Intelligent Text Processing*, (pp. 539-547).

Thornton, C. (2019, July 28). A General-Purpose Architecture for Text Mining-Domain-Independent Information Extraction Made Possible Through Natural Language Processing. *towards Data Science*.

Varsha Pande, D. (2016). Information Extraction Technique:A Review. *IOSR Journal of Computer Engineering*, 16-20.