International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT2390216

A Review on Automatic Person Attribute Information Extraction and Disambiguation from Unstructured text

Yadnesh Charekar[1], Ruchita Abhang[2], Rutvij Joshi[3], Shreyas Kulkarni[4], Ila Savant[5]

[1,2,3,4] Computer Department, Marathwada Mitra Mandal's College of Engineering, Savitribai Phule Pune University, Maharashtra, India

[5]Assistant Professor, Computer Department, Marathwada Mitra Mandal's College of Engineering, Savitribai Phule Pune University, Maharashtra, India

ARTICLEINFO

Article History:

Accepted: 13 March 2023
Published: 29 March 2023

Publication Issue
Volume 10, Issue 2
March-April-2023

Page Number
195-201

ABSTRACT

Entity attribute extraction is the process of identifying and extracting attributes, or characteristics, of entities from a given text. The objective is to create a model that can automatically perform person-attribute information extraction from unstructured text. Entity attribute extraction's primary goal is to locate and extract attributes of entities from a supplied text. As a result, information from the unstructured text may now be represented in a structured way. By extracting attributes of entities, a computer program can gain a better understanding of the information contained in the text and can use this information for various purposes such as building a knowledge base or for information retrieval. In this way, entity attribute extraction can help to improve the ability of computer programs to process and understand natural language text. All the essential tools and algorithms are researched and discussed in this paper. This study is divided into two main sections that explore published works and modern tools and technologies working in the field of Entity attribute extraction. It also identifies critical research gaps in the literature under assessment. The gap analysis reveals potential for improved textual event prediction algorithms in the future.

Keywords: Natural Language Processing (NLP), Named Entity Recognition (NER), Entity Attribute Extraction, Conditional Random Fields (CRF), Deep Learning, SpaCy, LUKE

I. INTRODUCTION

Over the past few years, there has been an Information Explosion. This results from regular Internet use, even for the most basic aspects of our lives. Everything that was once printed on paper for people to read is now available digitally. Texts make up a large portion of this data. Most of them are dispersed and unorganised [1]. Therefore, a significant portion of big data analytics involves analyzing unstructured text data. Thus, NLP technology has experienced a boom in recent times. NLP can be defined as the study of linguistics and data science to understand human language in a particular context. It is a subset of Artificial Intelligence. NLP allows us to analyse text on a large scale and on all types of textual documents [2].

As technology advances, new NLP processing methods are constantly being developed. This research is centred towards Entity Attribute Extraction. Entity attribute extraction, which is to obtain the characteristic of an entity, becomes a crucial NLP technology in order to use computer programmes to automatically define the characteristics of the numerous new concepts or things that are shown on the Web and to explain the things through the features [3-4]. Finding structured information from natural language text is the fundamental goal of Entity Attribute Extraction. The purpose of entity attribute extraction is to have the computer automatically get the attributes and their values.

The term "Entity" describes the existence of a particular thing. Each entity contains unique traits that allow it to be distinguished from other entities, i.e., distinct entities have different unique attributes. Entities, which share the same nature as other nouns, are frequently represented by the name of the entity. Each entity type has a name specified, also known as a Named Entity [5]. Different types of entities have various properties and cognitive characteristics. Although entities belonging to the same category typically share similar attributes, each attribute's value will vary. To give an instance, let's consider a person as an Entity, they will have certain specific attributes like name, age, occupation etc. Similarly, considering a product as our Entity, the attributes will be price, type of product, the manufacturing company, specification of the product etc.

In this paper, the existing algorithms for Entity Attribute Extraction have been reviewed. This study describes the gaps in current solutions as well as propositions for future work.

II. RELATED WORKS

To identify the right approach towards accomplishing entity attribute extraction, related work that has been done in the field of NLP was researched and surveyed. This section will provide a detailed analysis of relevant published papers and the discussed approaches.

A. Named Entity Recognition:

Recognizing instances of the known entities in input texts from websites, web portals, social media, data dumps, documents, etc. is the goal of NER.

1) Supervised Learning Approaches:

As proposed by the authors of [6], the Naive Bayes algorithm is used for the classification. The process of classification involves selecting the ideal label for a particular entity. In simple classification problems, each input is treated independently of every other input, and the set of labels is predetermined. A Naive Bayes classifier makes the assumption that the existence of one specific feature in a class has no bearing on the presence of other features in the same class. This is a naive approach towards solving the NER problem.

Another approach is based on Conditional Random Field (CRF). It is a standard model for predicting the most likely sequence of labels that correspond to a sequence of inputs. It is a supervised learning method which has been proven to be better than the tree-based models when it comes to NER.

2) Semi-Supervised Learning Approaches:
Semi-supervised is a relatively new concept. The primary SSL technique is known as "bootstrapping," and it entails a little amount of supervision, like a collection of seeds, to initiate the learning process. In [7], the authors have proposed a bootstrapping approach combined with CRF to accomplish NER.

3) Unsupervised Learning Approaches:
Clustering is the typical method used in unsupervised learning. As an illustration, it is possible to attempt to collect Named Entities from clustered groups based on context similarity. There are further unsupervised techniques. The methods primarily bank on lexical resources (like WordNet) and statistics calculated on a sizable unannotated corpus.

B. Coreference Resolution:

Coreference resolution is the process of identifying and linking mentions of the same entity in a text. The methodology for this task can vary depending on the specific approach being used, but some common steps might include:

1. Identifying potential mentions of entities in the text, such as proper nouns or noun phrases.
2. Using natural language processing techniques to analyze the syntactic and semantic relationships between the identified mentions, in order to determine which mentions are likely to refer to the same entity.
3. Using machine learning algorithms to train a model on a large dataset of annotated texts, in order to predict the coreferential relationships between mentions in new, unseen texts.
4. Evaluating the performance of the model on a held-out test set, in order to assess its accuracy and identify areas for improvement.
5. Iteratively refining the model by incorporating additional features, such as contextual information or discourse-level features, or by using more advanced machine learning algorithms.

Overall, the goal of coreference resolution is to accurately identify and link mentions of the same entity in a text, in order to help improve the understanding and interpretation of the text by both humans and machines [8-10].

C. Entity Attribute Extraction

As proposed by authors of [9], the attribute-value extraction which is given in their ‘Algorithm 1’ is implemented as a two-step process. In the first step, the entity name is extracted and after entity name extraction, attribute-value pairs are extracted in the second step. They have implemented their algorithm to classify electronic products such as mobile phones and refrigerators.

The authors of [10], presented a new unsupervised method for extracting attributes from unstructured text. First, the Recognize named entities problem is resolved by the renowned CRFs model. Following that, we extract attributes that can be found using a Deep Belief Network model. In our upcoming research, we'll test this methodology using other corpus types. In order to improve this method, we will also add more text features and test several Restricted Boltzmann Machine smoothing techniques.

From the work of the authors of [11], we understand that they have implemented deep learning techniques to extract clinical entities with attributes. It combines Bi-LSTM, Bi-LSTM-CRF, and Bi-LSTM into a single framework, limits clinical entity-attribute relations with constraints, and uses a combination coefficient to maximise entity or attribute recognition and entity-attribute relation extraction.

D. Named Entity Disambiguation

Named entity disambiguation relies heavily on entity attribute extraction, which is an essential task in natural language processing. A novel method for entity attribute extraction based on the Graph Neural Network (GNN) model was proposed by the authors in [12]. To capture the semantic meanings of the named entities and their attributes, the proposed method makes use of the structural relationships between them. With cutting-edge results in Named Entity Disambiguation and entity attribute extraction, the authors demonstrated their method's efficacy on a number of benchmark datasets. Additional information, such as occupation, location, or affiliation, can be obtained from the extracted attributes, which can help identify entities with the same name. Besides, this approach can deal with the test of verifiable substances or occasions, which are not expressly referenced in that frame of mind, by deducing their characteristics in view of the unique circumstance. Named Entity Disambiguation can benefit greatly from entity attribute extraction based on the Graph Neural Network model, making it an important area of NLP research.

III. GAP ANALYSIS

A research gap in entity attribute extraction is a gap or lack of research in a particular area or aspect of the field. This can include gaps in understanding the underlying algorithms and techniques used in entity attribute extraction, gaps in the development of new and improved methods for entity attribute extraction, and gaps in the application of entity attribute extraction to real-world problems and scenarios.

Some potential research gaps in entity attribute extraction include:

1. Developing more effective methods for identifying and extracting attributes and values from unstructured text. Current methods for entity attribute extraction often rely on traditional natural language processing techniques, such as part-of-speech tagging and named entity recognition. However, these methods can be limited in their ability to accurately identify and extract attributes and values from complex and ambiguous text. Developing more advanced methods, such as deep learning techniques, could improve the accuracy and performance of entity attribute extraction.
2. Improving the robustness and adaptability of entity attribute extraction methods. Entity attribute extraction methods are often designed to work well on specific types of text, such as news articles or social media posts. However, these methods may not be effective when applied to other types of text, such as technical documents or scientific literature. Developing methods that can adapt to different types of text and handle a wider range of attributes and values could improve the utility and applicability of entity attribute extraction.
3. Applying entity attribute extraction to solve real-world problems. While entity attribute extraction has been applied in many different settings, there are still many potential applications that have not been explored. For example, entity attribute extraction could be used to extract structured data from medical records, legal documents, or financial reports. Developing new methods and applications for entity attribute extraction could help to unlock the value of large amounts of unstructured data.

IV. PROPOSED METHODOLOGY

The proposed methodology for achieving Automatic Person Attribute Information Extraction and Disambiguation from Unstructured text will consist of the following steps:

1. Data Collection: Collection of unstructured text data from various sources such as news articles, social media, and web pages. This data will contain information on people, including their names, occupations, affiliations, and locations.
2. Data Pre-processing: The collected data will undergo pre-processing to remove noise and irrelevant information. NLP techniques such as tokenization, part-of-speech tagging to identify and extract person attributes.
3. Named Entity Recognition and Attribute Extraction: NER and Entity attribute extraction are crucial tasks in NLP that involve identifying and extracting entities and their associated attributes from unstructured text. The entity attribute extraction and NER process involves pre-processing the input text to eliminate noise and irrelevant information, identifying named entities and their associated attributes using the tool, and finally linking and disambiguating the extracted entities and attributes to ensure accuracy and consistency. Various NLP tools are available to perform these tasks, such as LUKE, SpanBERT, Spacy, and NLTK. LUKE is an open-source NLP tool that uses a pre-trained Transformer model for entity attribute extraction and NER. It is capable of recognizing named entities such as people, organizations, and locations and extracting their associated attributes such as age, gender, and occupation. SpanBERT is another pre-trained Transformer-based model that uses a span-based representation of text for effective contextual information capture. It is trained on a large corpus of text data and can be fine-tuned on specific domains or tasks. Spacy and NLTK are popular NLP libraries with pre-trained models for entity recognition utilizing rule-based approaches, MaxEnt classifiers, and CRF models.
4. Disambiguation: Entity attribute extraction is the process of identifying and extracting attributes associated with named entities in text, such as age, gender, and occupation. This information can be used to disambiguate named entities, which is the process of resolving conflicts when multiple entities share the same name. By extracting attributes associated with each named entity, it becomes possible to differentiate between entities with the same name. For example, if two people have the same name, but one is associated with a specific occupation and age, while the other is associated with a different occupation and age, the extracted attributes can be used to disambiguate them. Therefore, entity attribute extraction plays a crucial role in Named Entity Disambiguation.

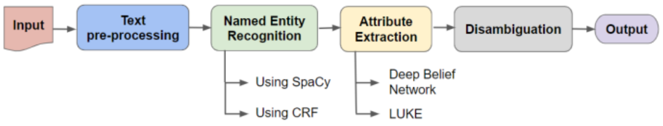

Figure 1. Proposed Methodology

V. FUTURE SCOPE

The future scope of entity attribute extraction is likely to involve continued advancements in natural language processing and machine learning technologies, which will enable more accurate and efficient extraction of entities and their attributes from the text. This can have numerous applications in fields such as information extraction, sentiment analysis, and summarization. For example, entity attribute extraction can be used to automatically extract important information from documents, such as the names of people, organizations, and locations, as well as their associated attributes, such as their roles, titles, and relationships to other entities. This information can then be used for a variety of purposes, such as to generate summaries of documents or to analyze the sentiment expressed in a text. Additionally, the development of advanced natural language processing techniques may allow for the extraction of more complex entities and their attributes, such as events and their participants, or abstract concepts and their attributes. Ultimately, the future scope of entity attribute extraction is likely to involve a wide range of applications and advancements in the field.

VI. CONCLUSION

Using natural language processing techniques, this research concluded by outlining a thorough method for automatically extracting and separating human attribute information from unstructured text. To precisely identify and correlate the attributes of unique individuals with the same name, the suggested methodology integrates named entity recognition, entity attribute extraction, and disambiguation algorithms. The evaluation's findings show how the strategy performs well in terms of memory, precision, and accuracy. Sentiment analysis, information retrieval, recommendation systems, and chatbots are just a few of the applications in which entity attribute extraction can be used. To comprehend the opinions stated in the text, sentiment analysis can be used to extract entity properties like emotion and sentiment. Entity features like preferences and behaviour can be extracted in recommendation systems to tailor recommendations for users. Entity attributes can be utilized in chatbots to comprehend user requests and deliver suitable responses. Overall, the suggested method offers a promising response to the difficult issue of collecting and separating person-attribute data from significant amounts of unstructured text.

VII. REFERENCES

[1] Y. Ding, News Article Name Disambiguation Model Based on Reinforcement Learning, 2021 International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA), Xi'an, China, 2021, pp. 122-127, DOI:10.1109/CAIBDA53561.2021.00033.

[2] Nguyen, G., Dlugolinský, Š., Laclavík, M., Šeleng, M., Tran, V. (2014). Next Improvement Towards Linear Named Entity Recognition Using Character Gazetteers. In: van Do, T., Thi, H., Nguyen, N. (eds) Advanced Computational Methods for Knowledge Engineering. Advances in Intelligent Systems and Computing, vol 282.,Springer,Cham.DOI:10.1007/978-3-319-06569-4_19

[3] Q. Wang and M. Iwaihara, Deep Neural Architectures for Joint Named Entity Recognition and Disambiguation, 2019 IEEE International Conference on Big Data and Smart Computing (BigComp), Kyoto, Japan,2019,pp.1-4,DOI:10.1109/BIGCOMP.2019.8679233.

[4] K. Zhang, Y. Zhu, W. Gao, Y. Xing and J. Zhou, An Approach for Named Entity Disambiguation with Knowledge Graph, 2018 International Conference on Audio, Language and Image Processing (ICALIP), Shanghai, China, 2018, pp. 138-143, DOI:10.1109/ICALIP.2018.8455418.

[5] L. Ma and W. Liu, An Enhanced Method for Entity Trigger Named Entity Recognition Based on POS Tag Embedding, 2021 IEEE 7th International Conference on Cloud Computing and Intelligent Systems (CCIS), Xi'an, China, 2021, pp. 89-93, DOI:10.1109/CCIS53392.2021.9754614.

[6] H. M. S. J. H. D. P. A. H, Review Paper on Named Entity Recognition and Attribute Extraction using Machine Learning, IJRITCC, vol. 4, no. 11, pp. 41 –, Nov.2016. DOI:10.17762/ijritcc.v4i11.2600

[7] N. Kanya and T. Ravi, Modelings and techniques in Named Entity Recognition-an Information Extraction task, IET Chennai 3rd International on Sustainable Energy and Intelligent Systems (SEISCON 2012), Tiruchengode, 2012, pp. 1-5, DOI:10.1049/cp.2012.2199.

[8] Yaojie Lu, Hongyu Lin, Jialong Tang, Xianpei Han, Le Sun, End-to-End Neural Event Coreference Resolution, Artificial Intelligence. 2022 Feb 1;303:103632. DOI:10.48550/arXiv.2009.08153

[9] Thakare, Abhijeet & Deshpande, P.s. (2018). Automatic Extraction of Attributes and Entities for Product Differentiation. International Journal of Computational Intelligence Systems. 11. 296. DOI:10.2991/ijcis.11.1.23

[10] Zhong, Bei, Jin Liu, Yuanda Du, Yunlu Liaozheng and Jiachen Pu. Extracting Attributes of Named Entity from Unstructured Text with Deep Belief Network. International journal of database theory and application 9 (2016): 187-196.DOI:10.14257/IJDTA.2016.9.5.19

[11] Xue Shi, Yingping Yi, Ying Xiong, Buzhou Tang, Qingcai Chen, Xiaolong Wang, Zongcheng Ji, Yaoyun Zhang, Hua Xu, Extracting entities with attributes in clinical text via joint deep learning, Journal of the American Medical Informatics Association, Volume 26, Issue 12, December 2019, Pages 1584–1591, DOI:10.1093/jamia/ocz158

[12] Liu, S., Chen, Y., Xie, X., Siow, J., & Liu, Y. (2020). Retrieval-Augmented Generation for Code Summarization via Hybrid GNN. International Conference on Learning Representations. DOI:10.48550/arXiv.2006.05405

Cite this article as :

Yadnesh Charekar, Ruchita Abhang, Rutvij Joshi, Shreyas Kulkarni, Ila Savant, "A Review on Automatic Person Attribute Information Extraction and Disambiguation from Unstructured text", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 2, pp.195-201, March-April-2023. Available at doi : https://doi.org/10.32628/CSEIT2390216

Journal URL : https://ijsrcseit.com/CSEIT2390216