International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 05 | May - 2022 Impact Factor: 7.185 ISSN: 2582-3930

INFORMATION EXTRACTION USING NATURAL LANGUAGE PROCESSING

Varun Reji , Srikanth Reddy N , Umar Shariff

Abstraction:

The expanding volume of data/Information creates a new challenge for Information Extraction techniques. The amount of unstructured data has increased in recent years. When we use this data in a clean and better way, we can extract wide variety of beneficial outcomes. Reduction of a text to its essential content, is a very complex problem which, despite the progress in the area thus far, poses many challenges to the scientific community. It is also a relevant application in today's information society given the enormous amount of production and processing of data, leading to exponential growth of textual information online and the need to momentarily assess the contents of text collections. Information Extraction using Natural Language Text is used to process, or extract and encode information from unstructured data and produce desired output data according to the application. This review article is about how Information Extraction is done using NLP (Natural language Processing)

Introduction

Information Extraction (IE) is a crucial cog in the field of Natural Language Processing (NLP) and linguistics. It's widely used for tasks such as Question Answering Systems, Machine Translation, Entity Extraction, Event Extraction, Named Entity Linking, Relation Extraction, etc. I think Information extraction can reduce human effort, reduce expenses, and make the process less error-prone and more efficient.

Information extraction (IE) is an important and growing field, in part because of the development of ubiquitous social media networking millions of people and producing huge collections of textual information, data in the world that needs to be collected, studied, and organized daily and Mined information is being used in a wide array of application areas from targeted marketing of products to intelligence gathering for military and security needs.

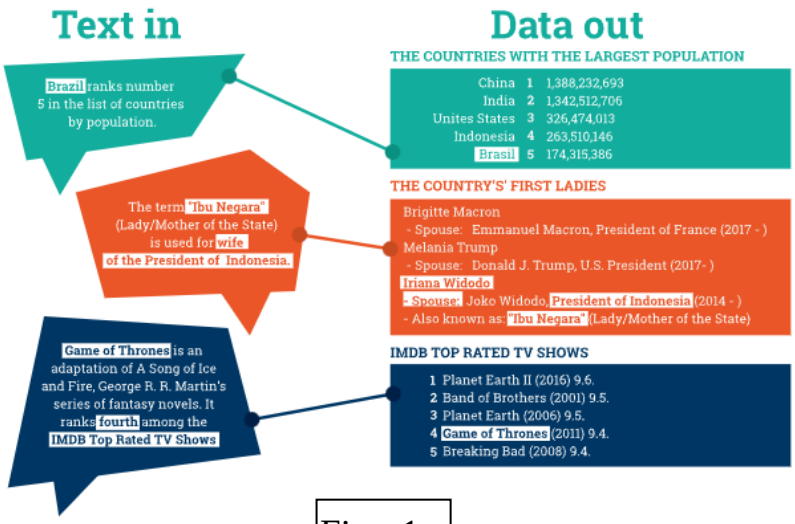

Fig - 1

IE has its roots in artificial intelligence fields including machine learning, logic and search algorithms, computational linguistics, and pattern recognition. Information Extraction from text data can be achieved by leveraging Deep Learning and NLP techniques like Name Entity Recognition. However, if we build one from scratch, we should decide the algorithm considering the type of data we're working on, such as invoices, medical reports, etc. This is to make sure the model is specific to a particular use case. To understand the mechanics

of Information Extraction NLP algorithms, we should understand the kind of data we are working on. This will help us to sort out the information we want to extract from the unstructured data. For example, for invoice related information, the algorithm should understand the invoice items, company name, billing address etc. While working on medical reports, it should identify and extract patient names, drug information, and other general reports. After curating the data, we'll then start applying the information extraction NLP techniques, to process and build models around the data. This article will delve into building Information Extraction algorithms on unstructured data using NLP techniques.

WHAT IS INFORMATION EXTRACTION AND HOW DOES IT WORK?

Information extraction is the process of extracting information from unstructured text data using NLP algorithms and converting them into editable and structured formant. Clever NLP techniques can be used to automate manual tasks.

Deep Learning and NLP techniques like Named Entity Recognition may be used to extract information from text input. Information Extraction System is used in a variety of NLP-based applications. For example, extracting relevant information from large collections of data/text like Wikipedia, Text Documents, real-time conversational AI systems like Chatbots, collecting weather reports for forecasting based on past records, fetching Patient's Medical records, Multilingual and Bilingual Information extraction.

We must first identify the type of data we are dealing with to comprehend the mechanics of Information Extraction using NLP techniques. This will assist us in separating the information we need from the unstructured data and to process it to structured and readily usable data.

Despite the mass availability of textual data, the complexity of Natural language makes extracting usable information from it extremely challenging. Regardless of how difficult the Information Extraction process is, practically all IE systems have a pipeline with certain similar phases.

NLP primarily comprises of natural language understanding (human to machine) and Natural Language Generation (machine to human). In recent years there has been a surge in unstructured data in the form of text, videos, audio, and photos. NLU aids in extracting valuable information from text such as social media data, customer surveys, and complaints.

Detailed Information

In this section, we will look at some of the ***General techniques, Linguistic Knowledge in Natural Language Processing*** and ***NLP techniques for Extracting Information***. Information Extraction systems takes natural language as its input and produces a structured information based on certain criteria. The IE systems can be applied to a wide range of document sources. For example, emails, reports, presentation, web pages, scientific papers, etc.

GENERAL TECHNIQUES

Information Extraction Techniques using Natural Language Processing uses some common steps involved in extraction of structured information from unstructured data.

1) ***Initial Processing:*** In the first step, the text is divided into fragments such as segments or tokens. This process can be achieved by using tokenizers, text zoners, segmenters, or splitters. With the help of morphological analysis, includes part-of-speech tagging and the identification of phrasal units (noun or verb phrases) are identified.

2) ***Proper names Identification***: Identification of various classes of proper names, such as names of people or organization, dates, monetary amounts, place, addresses, etc, is one of the most important operations in the chain of IE Operation.

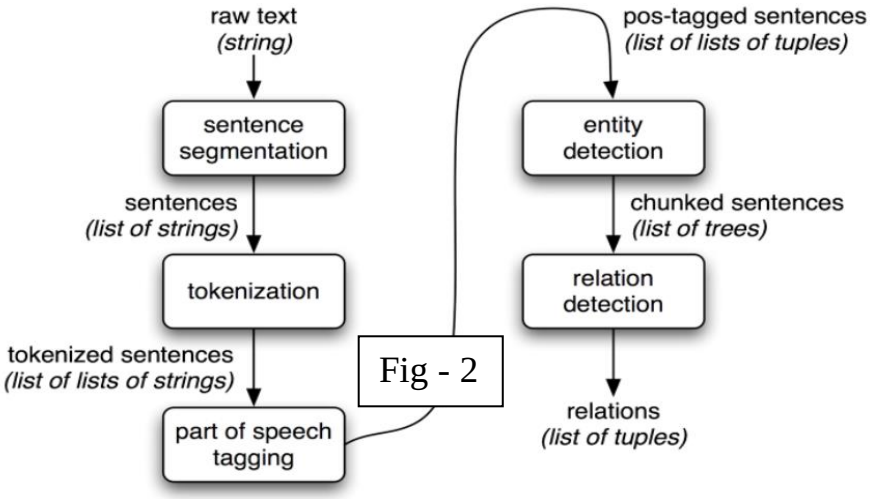

Fig - 2

3) ***Parsing:*** The syntactic analysis of the sentences in the documents is performed to identify the noun groups and verb groups, which are used at the pattern matching stage. Parsing is not an easy task, therefore expensive computations are involved. Partial or shallow parsing is used instead, and it creates partial, not overlapping syntactic fragments with a higher level of confidence.

4) ***Extraction of events and relations***: The major stage in the information extraction process is the extraction of events and relations. This is done by matching the text against extraction rules. The text is compared to certain patterns, and any match is discovered, the text is labelled and retrieved later.

5) ***Coreferences*** **or** ***Anaphora Resolution***: The MUC-6 introduced the coreference task, which is the resolution of anaphora's in texts. Several types of coreference are recognized, and the most common are pronominal and proper names coreference. The step where noun phrases are decides if related to the same entity or not is called Coreference or Anaphora Resolution.

6) ***Output results Generation***: The output results generation stage transforms the extracted structures into the output templates according to the format specified by the client. This includes variety of normalization processes.

LINGUISTIC KNOWLEDGE IN NLP

Machine Learning, Deep Learning, Statistics, Grammar and Formal Languages, Regular Expressions are some disciplines in Natural Language Processing. The different levels of common Natural Language Processing tasks can be built one upon another. This is represented using a pyramid.

1) ***Morphology***: This is responsible for making up the sentence using words and deals with the formation with the help of prefixes/suffixes, Gender detection, word inflection, spellchecking, etc.

2) ***Syntax***: Syntax is responsible for making a proper word construction and determines the underlying structure of a sentence. It basically referred as grammar. This includes part-of-speech tagging, building Dependency and Syntax trees.

3) ***Semantics***: This is where the actual natural language understanding is delt with, which includes, Name Entity Extraction, Relational Extraction and many more. Semantics usually works on sentences, where a sentence is a sequence of words usually with some added semantics (like sense, role) attached.

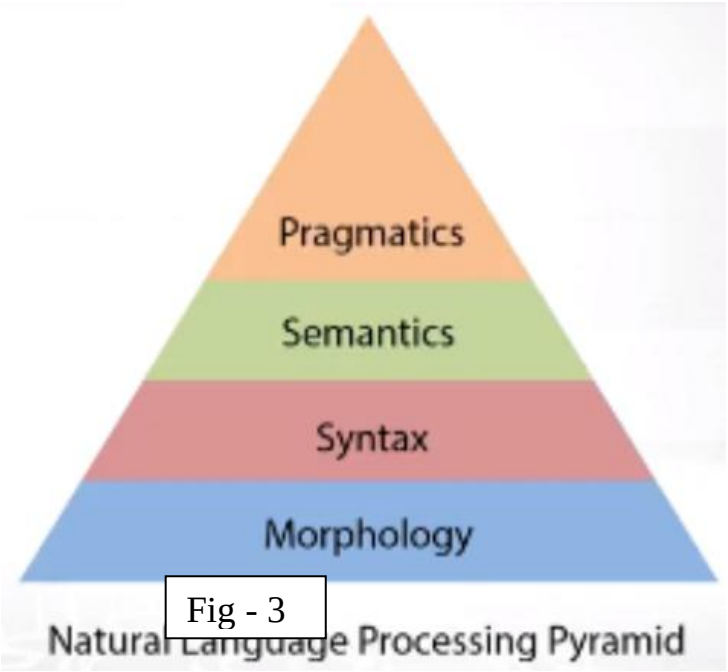

Fig - 3
Natural Language Processing Pyramid

4) ***Pragmatics***: This stage is responsible for find out the underlying threads, topics, and references. Topic segmentation, Lexical chains, Summarization are some of the discourse tasks.

NLP TECHNIQUES FOR EXTRACTING INFORMATION

The field of artificial intelligence has always envisioned machines being able to mimic the functioning and abilities of the human mind. Natural language processing (NLP) primarily comprises of natural language understanding (human to machine) and natural language generation (machine to human). This includes 5 common techniques –

1) ***Name Entity Recognition***: The most useful techniques in NLP used for pre-processing task. Identification of key information in text such as people, location, organization, dates, etc. This is based on grammar and supervised models.

2) ***Text Summarization***: Text summarization uses algorithms to generate fresh text that conveys the crux of the original text. LexRank, TextRank, and Latent Semantic Analysis are some examples of algorithms that can be used for text summarization.

Fig: 4

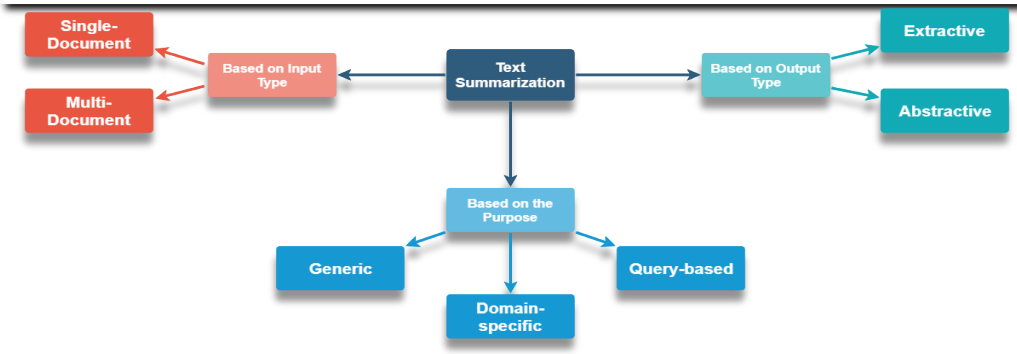

Fig: 5

3) ***Sentiment Analysis***: Sentiment analysis is used to find the sentiment of a text. In this case, we can use sentiment analysis to separate the positive and negative parts of the review. Sentiment analysis can be done using supervised and unsupervised techniques. The most popular supervised technique is Naive Bayes, but other techniques like random forest or gradient boosting can also be used.

4) ***Aspect Mining***: Aspect mining identifies the different aspects in the text, and sentiment analysis conveys the complete intent of the text.

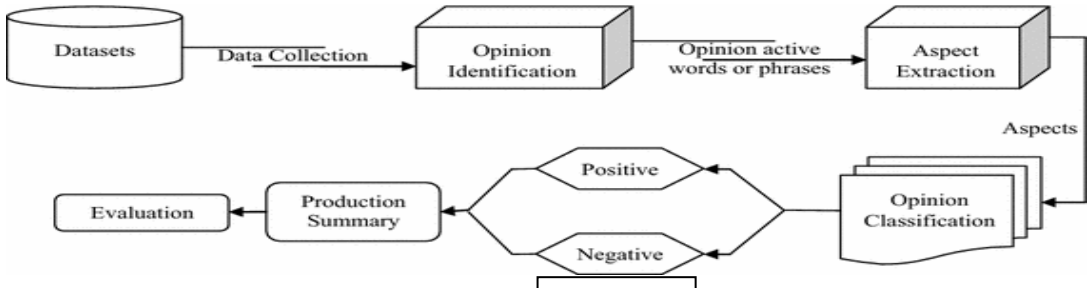

Fig: 6

5) ***Topic Modelling***: This includes Latent Sematic Analysis, Probabilistic Latent Semantic Analysis, Latent Dirichlet Allocation and Correlation Topic Model. Latent Dirichlet allocation (LDA) is a popular method for topic modelling. It uses the input text and the expected number of topics to identify the common words across two topics.

Conclusion

. The massive expanding volume of data/Information opens up to new possibilities and creates a new challenge for Information Extraction techniques. Information Extraction seems to be the best technique for extracting the text using NLP. Once the information is extracted from unstructured text using these methods, it can be directly consumed using machine learning models (Deep Learning and NLP Techniques) to enhance their accuracy and performance. Natural language processing (NLP) includes name entity recognition, text summarization, sentiment analysis, aspect mining, Latent Dirichlet allocation (LDA), and correlation topic model. These techniques are used for pre-processing tasks and for generating fresh text that conveys the crux of the original text.

References

[1] Turmo, *Jordi, Alicia Ageno, and Neus Catala. "Adaptive information extraction." ACM Computing Surveys (CSUR) 38, no. 2 (2006): 4-es.*

[2]https://towardsdatascience.com/linguistic-knowledge-in-natural-language-processing-332630f43ce1

[3] https://blog.aureusanalytics.com/blog/5-natural-language-processing-techniques-for-extracting-information

[4] Fig 1: https://www.ontotext.com/wp-content/uploads/2017/02/Text-in_Data-out.svg

[5]Fig2**:** https://media-exp1.licdn.com/dms/image/C5612AQGxsNqy6VbslQ/article-inline_image-shrink_1000_1488/0/1636908965632?e=1658361600&v=beta&t=EOC56kFTsVB7PuiGZdhreLtDXFaFi1_ekD4V_ONbva0

[6] Fig 3: https://miro.medium.com/max/700/1*Qtk5pN8n_BcYUsosrKFrFg.png

[7] Fig 4: https://aiimi.imgix.net/assets/images/Named-entity-recognition-aiimi-1.png?auto=format%2Ccompress&domain=aiimi.imgix.net&ixlib=php-3.3.0

[8] Fig 5: https://devopedia.org/images/article/261/5116.1582303416.png

[9] Fig 6: https://devopedia.org/images/article/234/8727.1572836738.gif