International Conference on Communication and Information Processing
(ICCIP-2019)
Available on: Elsevier-SSRN

A Review for Semantic Analysis and Text Document Annotation Using Natural Language Processing Techniques

Nikita Pande[a], Mandar Karyakarte[b]

[a]*Post Graduate student, Computer Engineering, VIIT, Pune-411 048, India*
[b]*Assistant Professor, Computer Engineering, VIIT, Pune-411 048, India*

Abstract

In today's fast-growing world with rapid change in technology, everyone wants to read out the main part of the document or website in no time, with a certainty of an event occurring or not. However annotating text manually by domain experts, for example cancer researchers or medical practitioner becomes a challenge as it requires qualified experts, also the process of annotating data manually is time consuming. A technique of syntactic analysis of text which process a logical form S-V-O triples for each sentence is used. In the past years, natural language processing and text mining becomes popular as it deals with text whose purpose is to communicate actual information and opinion. Using Natural Language Processing (NLP) techniques and Text Mining will increase the annotator productivity. There are lesser known experiments has been made in the field of uncertainty detection. With fast growing world there is lot of scope in the various fields where uncertainty play major role in deciding the probability of uncertain event. However, syntactic analysis alone will not give desired results. Hence, it is required to use different techniques for the extraction of important information on the basis of uncertainty of verbs and highlight the sentence.

Keywords - Annotation, NLP techniques, Text Mining, Uncertainty of verbs.

© 2019 – Authors

1. Introduction

Today technology has changed in a vast manner and every change in technology has made an impact on quality of human life. Among such, is availability and processing of digital data which are present over multiple sources like handheld devices, mobile apps, web apps, websites etc. The advances in digital data processing have evolved and have put forward multiple ways to read those available data rapidly in the condensed form. To make some

* Corresponding author. Tel.: +91-976-625-0550.
E-mail address: nikitapande128@gmail.com.

Electronic copy available at: https://ssrn.com/abstract=3418747

events or news eye catchy, different algorithm are used to highlight the content so that people can read those data easily, this is termed as Annotation. For example, "Cancer is very dangerous disease" has been highlighted using some colour.

Annotation of text refers to highlighting or underlining and it may include comments, tags, and notes. Text annotation with highlight will provide most important sentence and information quickly. Ismail et al., 2015 described annotation which provides the connection between topics and helps to get familiar with the content as well. The process of annotating text manually is time consuming and costly. Annotation of text will give important context of the document to the reader. The annotation of text may be carried out using different approaches like extraction of keywords having highest weights, keywords having lowest weights, etc described by Ismail et al., 2015. Researchers apply semantic analysis during the annotation procedure as semantic techniques yield solutions for extraction of entities, mapping amongst the entities and information retrieval challenges.

Earlier annotation of text was a manual process which used to take ample amount of time. While annotating the text user reads each line of a given data and tries to understand it. Then user may circle or underline the important or unfamiliar words or concepts as per his knowledge or requirements. this combined process gives an extraction of important sentences among the presented data which can be read quickly when needed. But user cannot annotate large number of documents in short period of time. Document interpretation done by accurate annotation of text which may require a knowledgeable reader. Therefore, extraction and annotation of text automatically for a large document becomes a feasible alternative.

Text Mining is useful to extract the important information from the text document. Text Mining is used for different processes like information extraction, topic detection, summarization, annotation, clustering and provides relevant information using information retrieval techniques. Natural Language Processing supports gathering reliable information from texts. Also, information techniques should be used for extracting relevant information from a repository or collected resources which provides information.

Natural Language Processing (NLP) can be used for a technique called Part-of-speech (POS) tagging for annotation in text which is corresponds to the part of speech. In pre-processing of text POS having major contribution. POS tagging can be done using different methods and algorithm. NLP libraries are available to perform POS tagging using python code which was implemented by Dudhabaware and Madankar, 2014. Semantic information of sentences can be obtained by using one more model called Sentence-Level Semantic Graph Model (SLSGM). Semantic analysis is used for the calculation of related values between sentences. For both internal and external information PageRank graph algorithm will give more feasible and effective results given by Han et al., 2016. NLP features and text analytics function can be enhanced, automate and accelerate by using Machine Learning. POS tagging, Sentiment analysis, entities are identified using statistical techniques of Machine Learning (ML) in NLP. Many researchers have found Support Vector Machine Algorithm as an effective ML algorithm.

Huge number of documents is available and additional documents are generated at rapid pace, to highlight the important text a system is required which will be able to annotate the text and extract important sentences using NLP.

2. Literature Review

Text mining provides wide range of knowledge sources in the field of biomedicine which helps to cancer researchers and oncologist. Baker 2018, proposed semantic text classification using two NLP classification task and datasets. The first task was classification according to the Hallmarks of Cancer (HoC), which provides the text mining of scientific literature that gives the explanation of the process through which the cancer starts and spread in the body. The second task provides the chemical routes exposures into the body that may lead carcinogens exposure. Conventional pipeline NLP approach is used for both the task also deep learning methods are used. Ismail et al. 2015, had suggested different methods used for the text document annotation like Keyword Based Annotation (KWBA) and Ontology Based Text Annotation (OBA). CAT tool, Word Sense Disambiguation (WSD) techniques were used to analyze text document and provides semantic information of the document automatically.

Above papers provides techniques for the semantic text classification of biomedical data and ontology-based data. This technique helps in annotating complex text related to biomedicine and cancer and also provides

Electronic copy available at: https://ssrn.com/abstract=3418747

annotation methods.

Stenetorp et al. 2012, introduced Brat rapid annotation tool which was a web-based tool used for text annotation. Natural Language Processing (NLP) had used in Brat which provides rich structured annotation for various NLP tasks. The client-server architecture was used for the implementation of Brat which provides the communication over HTTP using JavaScript Object Notation (JSON). A stateless server back-end is used in Brat which is implemented in Python which supports Common Gateway Interface (CGI) and FastCGI protocols. Machine Learning algorithm implemented for an experimental demonstration, reduced time for by 30% for type selection and by 15% above all annotation time for a multi-type entity.

Forbes et al. 2018, introduced Text Annotation Graphs, or TAG a web-based text annotation software tool. Complex relationship between words and word phrases was provided by TAG. The motivation behind TAG was BRAT annotation tool. BRAT does not provide the capability to produce link between link.

Subject-Object-Predicate were extracted from individual sentences to form a semantic graph of the original document. Leskovec et al. 2005, implemented Support Vector Machines (SVM) learning algorithm to train a classifier which identify SOP triples from the semantic graph which belongs to the summary. Automatic extraction of summaries from text document was achieved using this classifier. DUC 2002 and CAST datasets generates statistically significant improvements which included semantic properties and topological graph properties. Semantic graph attributes alone would not give good results.

To determine the candidates of cancer driver mutations list of known cancer can be used. Park et al. 2012, presented a system CaGe which was a web-accessible cancer genome annotation system. Reported cancer gene and cancer annotation databases were constructed from public cancer genomic databases. A gene ID database was also constructed to allow input of gene list from various input format. Output files like SIFT or PolyPhen were handled as input files for NGS-based cancer genomics flows.

Mohanty et al. 2007, proposed a method to generate the semantics of the source language sentences which were captured by building an Interlingua based machine translation system in the form of Universal Networking Language (UNL) graphs. Semantically relatable sequences (SRS) were formed by recognizing conceptual arguments. The right level of granularity and expressive power was provided by UNL representation.

Mohanty et al. 2008, implemented semantic representation by converting machine translation source language sentences. NLP resources like the WordNet and OALD and NLP tools like the parser were used for the computation of SRS. FrameNet corpus was used to generate Gold Standard SRSs which takes adjectives, verbs, nouns as targets. It yields 82% accuracy of SRS identification for 92,310 total numbers of sentences.

Hasan et al. 2018, proposed different techniques for the identification of the core content of a text particular number of key terms were extracted. Pre-processing and post-processing were the two steps used to build the model. Firstly, the document was divided into sentences and the stop words were removed by the model using a list of predefined stop words. Stemming was done by porter stemming method. For selected datasets Latent Semantic Analysis contributed by generating accuracy of 77.6% precision and 84.3% recall.

Eliminating redundant information is the major issue of text summarization. Han et al. 2016, proposed a method in which semantic information of sentence was given by Sentence-Level Semantic Graph Model (SLSGM). Sentences were considered as vertexes and the edges represent semantic relationship between the sentences. The relevance values between the sentences were calculated using semantic analysis which was reflected as a weight of edges. PageRank graph ranking algorithm was used to calculate sentence value. DUC 2004 dataset was used for the comparison of various models.

Dali et al. 2009, implemented a system where the list of facts associated with the answers of the questions were described by subject-verb-object triplets. Penn Treebank parse tree was formed for each sentence of document and subject-verb-object extracted automatically. Coreference resolution was used for named entity extraction. For the successful linguistic analysis, it required predefined grammatical structure of the question.

Temizer and Diri 2012, extracted Subject-Object-Verb automatically from Turkish documents using three main steps: morphological analysis, dependency analysis and triplet extraction Pre-processing of text was achieved by Java Standard Edition string processing facilities. Zemberek API was used for the morphological analysis of words. Rule based algorithm was used for the dependency analysis of sentences.

Rusu et al. 2007, presented an approach for generating a parse tree from the sentences, four different syntactic parsers for English were used. Parser dependent techniques were used for the extraction of triplets from the parse

Electronic copy available at: https://ssrn.com/abstract=3418747

tree. The parsers namely Stanford Parser, OpenNLP, Link Parser and Minipar were used for generating parse tree. A treebank structure was generated by OpenNLP and Stanford Parser. Link Parser produce result based on Link grammar. Triplet extraction algorithm was used which was given by Minipar to obtain parse tree. 110 triples were generated after parsing the sentences in 271 seconds

Rusu et al. 2009, proposed a technique for generating semantic graph in form of directed graph, was obtained based on the semantic representation of text. From the given document pronouns were identified using pronominal anaphora resolution heuristic. Scores were given to the pronoun possible candidates that could be able to replace it. Pronoun having highest score was selected for the replacement. Sentence complexity and document size mainly affects the runtime.

Dudhabaware and Madankar 2014, provided a review on NLP task with different application areas. The comparison is carried out to provide better NLP task for pre-processing of search keywords. Different NLP tasks were implemented using different pre-processing steps such as Coreference Resolution of text, Named Entity Recognition (NER), Stemming, Sentiment Analysis, etc. POS tagging and chunking gives more desirable result for pre-processing the keywords.

Todorovic et al. 2008, proposed Hidden Markov Method, a supervised method used for the implementation of NER which classify different entities from text into classes. Two different Hidden Markov Model implemented which provides surrounding words context to classify the named entity of current word.

Luo et al. 2012, implemented Conditional Random Field, a machine learning algorithm where semantic information could be used by CRFs model to present domain ontology features. It provides rich overlapping features by combining internal and external features to form compound feature. Sequence data was labelled and segmented using CRFs. It achieves overall F-measure around 87.16%. But the review text was complicated and Domain ontology was imperfect, the product named entity recognition may not provide better result as compared to traditional named entity recognition.

Zhonglin et al. 2016, proposed some methods based on dictionaries and traditional statistical machine learning which were combine to achieve POS tagging. For testing the data People's Daily corpus was used and POS tagging achieve 95.80% of accurate rate. SWJTU segmentation dictionary was used for POS tagging. This approach was used to solve three problems like ignoring the context information of words, POS tag errors which are present in the dictionary and insufficiency of the training corpus in statistical machine learning method.

Kerdjoudj and Cure 2015, presented a solution to qualify uncertainty of extracted text document. From linguistic analysis uncertainty of the extracted statement could be determine. The information of web document may contain uncertainty. Resource Description Framework (RDF) graph model was used for the representation of the information using SPARQL. RDF triples were used for the representation of different patterns of uncertainty.

Jean et al. 2016, has proposed a vector-based representation which was used based on statistical analysis of different lexical and syntactic features for the characterization of sentences. Extraction of lexical and syntactical features was carried out by using three corpora: Bioscope, WikiWeasel and SFU which provides set of annotated sentences. Bioscope and WikiWeasel generates F-measure 71.2%.

From above literature review, it is observed that there are different approaches used for annotating a text document like pre-processing of data, semantic analysis, syntactic analysis, statistical machine learning algorithm etc. These methods will help in extracting keywords, generating subject-verb-object from the sentences which was used for the highlighting of the important keywords. Each of these methods do have their advantages and disadvantages like BRAT does not provide the capability to produce link between link and triple extraction algorithm used to construct parse tree. But for complex sentence the complexity of parse tree is also increases. Annotating the text sentences is incomplete, if the mapping amongst the different entities, different forms of same entity is not established. The work presented by authors' viz. Park et al. 2012, Leskovec et al. 2005, Rusu et al. 2009 has a limitation to establish mapping amongst different morphs of same entity. The proposed work intends towards establishing such mappings, build context around entities and annotate the text document similar to manual annotation. The work shall be further enhanced for using combination of techniques used by the other researchers for annotating uncertainty in the text sentences.

Also, it is clear from the review that any solution to achieve the mentioned objectives, fundamental step is building parse tree based on NLP of text sentences. The next section presents building parse tree and how their

Electronic copy available at: https://ssrn.com/abstract=3418747

complexity does vary with respect to the size of input text data.

3. Computation Experiment

NLP techniques are used for the pre-processing of the text document. Using different libraries output was obtained. Tokenization of text is done by breaking a text into individual words. For parsing using NLP parser, these tokens are given as an input for analysis. POS tagging is done for marking up a part of speech tag with a word in the text. Parse tree can be built by using POS tags which can be further useful to form Named Entity Relations. NLTK Treebank dataset is used for POS tagging which gives better results.
For example: "The little yellow dog barked at the cat."
Here, sentence (S) is represented by two children noun phrase (NP) by the parser as a parse tree. "the" is a determiner (DT), "little" and "yellow" are tag as adjectives (JJ), "dog" and "cat" are singular noun (NN).

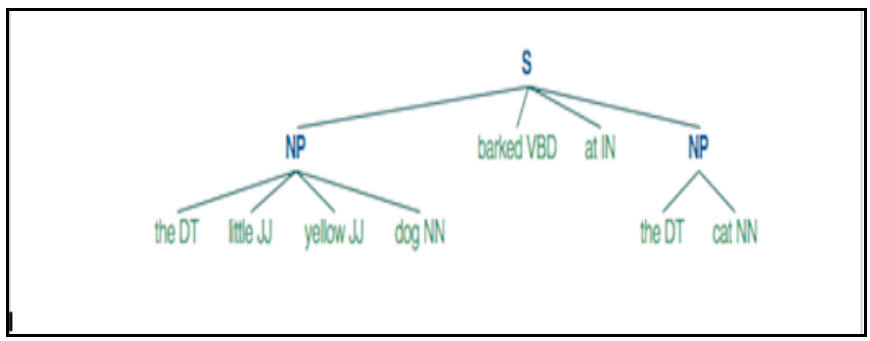

Fig. 1. Parse tree for single sentence.

For large chunk of data from website it will generate a complex parse tree which is hard to decode and chooses the relevant part of speech for annotating. Also, the processing time for POS Tagger is bit high as it parses each word and tag them to different part of speech.

Syntactic analysis and Semantic analysis mainly used for the better understanding of natural language. Semantic structure is assigned to text using syntactic analysis.
For example: "The Cat (noun-phrase) went away (verb-phrase).", where a noun-phrase is a Subject and a verb-phrase is a Predicate.

Subject-Verb-Object are extracted individually from each sentence using syntactic analysis to form logical triples which provides important information related to the text. This will give required output for annotating and highlighting the sentences. But alone syntactic analysis is not sufficient to get better output we need to implement different NLP techniques. Hence, on the basis of uncertainty of verbs we can be able to get more efficient output.

4. Conclusion and Future Scope:

In this paper, a review about the work done is carried out which provides different techniques used for annotation of text. Pre-processing of document text has been accomplished by using Natural Language Pre-processing. Different NLP task are performed for the implementation of pre-processing steps like tokenization, POS tagging. Syntactic analysis done to form logical form SVO triples which provides important information with respect to the sentence. The other text annotation techniques such as ML, KWBA, OBA, WSD are used to analyse text document and provides semantic information of the document automatically. There are different web-based tools like Brat, CaGe used for annotation of text. Also, uncertainty play major role in deciding the importance of the verbs. So, in future Annotating web-based text on the basis of uncertainty will acquire more efficient result in extracting key phrases and sentences.

Electronic copy available at: https://ssrn.com/abstract=3418747

References

[1] Baker, S., 2018. Semantic text classification for cancer text mining (Doctoral thesis).
[2] Ismail, I., Gad, W., Hamdy, M., Bahnsy, K., 2015. Text document annotation methods: Stat of art, IEEE Seventh International
[3] Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, pp. 634-640.
[4] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Anani-adou, S., Tsujii, J., 2012. BRAT: A web-based tool for NLP-assisted text annotation, In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 102–107.
[5] Forbes, A. G., Lee, K., Hahn-Powell, G., Valenzuela-Escarcega, M. A., Surdeanu, M., 2018. Text Annotation Graphs: Annotating Complex Natural Language Phenomena, In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18), Miyazaki, Japan. European Language Resources Association (ELRA).
[6] Leskovec, J., Milic-Frayling, N., Grobelnik, M., 2005. Extracting Summary Sentences Based on the Document Semantic Graph, MSR-TR-2005-07.
[7] Park, Y. K., Kang, T. W., Baek, S. J., Kim, K. I., Kim, S. Y., Lee, D., & Kim, Y. S., 2012. CaGe: A Web-Based Cancer Gene Annotation System for Cancer Genomics, Genomics & informatics, 10(1), 33-9.
[8] Mohanty, R.K., Prasad, M.K., Narayanaswamy, L., Bhattacharyya, P., 2007. Semantically Relatable Sequences in the Context of Interlingua Based Machine Translation, In the proceedings of the 5th International Conference on Natural Language Processing, Hyderabad, pp. 1-8, 2007.
[9] Mohanty, R., Limaye, S., Prasad, M.K., Bhattacharyya, P., 2008. Semantic Graph from English Sentences, Proceedings of ICON 6th International Conference on Natural Language Processing Macmillan Publishers, India.
[10] H. M. M. Hasan, F. Sanyal and D. Chaki, 2018. A Novel Approach to Extract Important Keywords from Documents Applying Latent Semantic Analysis, 10th International Conference on Knowledge and Smart Technology (KST), Chiang Mai, 2018, pp. 117-122.
[11] Han, X., Lv, T., Jiang, Q., Wang, X., Wang, C., 2016. Text summarization using sentence-level semantic graph model, 4th International Conference on Cloud Computing and Intelligence Systems (CCIS), pp 171–176.
[12] Dali, L., Rusu, D., Fortuna,B., Mladenić, D., Grobelnik, M., 2009. Question Answering Based on Semantic Graphs, Workshop on Semantic Search at WWW2009, Madrid, Spain.
[13] Temizer, M., Diri, B., 2012. Automatic subject-object-verb relation extraction, Proc. 2012 Int. Symp. Innovations in Intelligent Systems and Applications, pp. 1–4.
[14] Rusu, D., Dali, L., Fortuna, B., Grobelnik, M., Mladeni, D., 2007. Triplet extraction from sentences, In Proceedings of the 10th international multiconference information society vol. A, pp. 218–222.
[15] Rusu, D., Fortuna, B., Mladenić, D., Grobelnik, M., Sipoš, R., 2009. Visual analysis of documents with semantic graphs, Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration, p.66-73.
[16] Dudhabaware, R.S., Madankar, M.S., 2014. Review on natural language processing tasks for text documents, IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, pp. 1-5.
[17] Todorovic, B.T., Rancic, S.R., Markovic, I.M., Mulalic, E.H., Ilic, V.M., 2008. Named Entity Recognition and Classification using Context Hidden Markov Model, 9th Symposium on Neural Network Application in Electrical Engineering, NEUREL, pp. 43-46.
[18] Luo, F., Fang, P., Qiu, Q., Xiao, H., 2012. Features Induction for Product Named Entity Recognition with CRFs, Proceedings of the IEEE 16th International Conference on Computer Supported Cooperative Work in Design, pp. 491-496.
[19] Ye, Z., Jia, Z., Huang, J., Yin, H., 2016. Part-of-speech tagging based on dictionary and statistical machine learning, 35th Chinese Control Conference (CCC), Chengdu, pp. 6993-6998.
[20] Kerdjoudj, F., Curé, O. 2015. Evaluating uncertainty in textual document, Proceedings of the Eleventh International Workshop on Uncertainty Reasoning for the Semantic Web (p. 1). Bethlehem, PA, USA: Springer.
[21] Jean, P.A., Harispe, S., Ranwez, S., Bellot, P., Montmain, J., 2016. Uncertainty detection in natural language: A probabilistic model, International Conference on Web Intelligence Mining and Semantics, pp. 10:1-10:10.

Electronic copy available at: https://ssrn.com/abstract=3418747