Event Extraction from News Articles

Siddhant Agarwal
IIT2020228

***Abstract*—Event extraction from news articles is a commonly required prerequisite for various tasks, such as article summarization, article clustering, and news aggregation. Due to the lack of universally applicable and publicly available methods tailored to news datasets, many researchers redundantly implement event extraction methods for their own projects. The journalistic 5W1H questions are capable of describing the main event of an article, i.e., by answering who did what, when, where, why, and how. We provide an in-depth description of an improved version of Giveme5W1H, a system that uses syntactic and domain-specific rules to automatically extract the relevant phrases from English news articles to provide answers to these 5W1H questions. Given the answers to these questions, the system determines an article's main event.**

***Index Terms*—News Event Detection, 5W1H Extraction, 5W1H Question Answering, Reporter's Questions, Journalist's Questions, 5W QA**

I. Introduction

The 5W1H technique is a powerful and versatile tool used for problem-solving, decision-making, and gathering information. The technique is based on six key questions: who, what, where, when, why, and how. By answering these questions, a complete and detailed understanding of a problem can be gained, which can help in finding solutions and making informed decisions.

The technique is widely used in various fields, including journalism, investigations, research, and project management. In journalism, the 5W1H technique is used to gather and report news stories. In investigations, it helps to gather evidence and identify suspects. In research, it helps to formulate research questions and design research studies. In project management, it helps to identify project goals, constraints, and risks.

The 5W1H technique is simple and easy to use, yet powerful and effective. It can be used by individuals and teams to ensure that all aspects of a problem are considered and addressed. The technique encourages clear and concise communication, and helps to prevent misunderstandings and confusion. It also helps to identify information gaps and areas that require further investigation.

Overall, the 5W1H technique is an essential tool for anyone who needs to gather information, analyze a situation, or solve a problem. Its simplicity and versatility make it a valuable tool in a wide range of applications.

II. Literature Survey

A. Paper 1

Author- Felix Hamborg, Corinna Breitinger, Bela Gipp

Title- Giveme5W1H: A Universal System for Extracting Main Events from News Articles

Methodology- Giveme5W1H accepts the complete text of a news article, including the headline, lead paragraph, and body text, as input. During preprocessing, they use Stanford CoreNLP for sentence splitting, tokenization, lemmatization, POS-tagging, full parsing, NER (with Stanford NER's seven-class model), and pronominal and nominal coreference resolution. Once the initial preprocessing is completed, all Named Entities (NEs) in the text are converted into their canonical form.

Then it performs four independent extraction chains to retrieve the article's main event: (1) the action chain extracts phrases for the 'who' and 'what' questions, (2) environment for 'when' and 'where', (3) cause for 'why', and (4) method for 'how'.

The final step is to select the most appropriate candidate for each of the 5W1H questions. This process involves two sub-tasks: individual scoring of candidates for each question, and combined scoring that takes into account the properties of candidates for other questions. To score candidates for each question, we use a weighted sum of n scoring factors. For the combined scoring, we adjust the scores of candidates for one question based on the properties of candidates for other questions.

Performance- The dataset was divided into 6 categories of Bus, Ent, Pol, Spo and Tec. For Bus average accuracy for all 5W1H was .72, for Ent avg was .72, for Pol avg was .75, for Spo avg was .73, for Tec avg was .71 and for all the categories included average accuracy was .73.

Dataset- For the English language, the authors used the New York Times Annotated Corpus (NYTAC) as the main dataset. The NYTAC consists of over 1.8 million articles from The New York Times spanning from 1987 to 2007. The authors also used the Reuters Corpus Volume 1 (RCV1) and the English Gigaword corpus as supplementary datasets for their evaluation.

For the Spanish language, the authors used the Europarl Corpus, which contains transcripts of speeches from the European Parliament, as the main dataset. They also used the Spanish Gigaword corpus as a supplementary dataset.

For the Chinese language, the authors used the Sina News corpus, which consists of news articles from the Sina News website, as the main dataset. They also used the Chinese Gigaword corpus as a supplementary dataset.

Year- 2019

B. Paper 2

Author- Jin Zhang, Tao Li, and Wei Xu

Title- Event Detection and Clustering Using Maximum Entropy Based on Semantic and Temporal Features

Methodology- The main objective of the paper is to propose a novel approach for event detection and clustering by combining both semantic and temporal features. The proposed method uses a maximum entropy model to classify the news articles into different events and cluster them accordingly.

The paper uses a dataset of news articles collected from various sources such as Reuters and New York Times, covering a wide range of topics such as sports, politics, and entertainment. The dataset consists of around 10,000 news articles, which are manually labeled into different events by human annotators.

The proposed method first extracts the semantic features of the news articles using a combination of N-gram and latent semantic analysis (LSA) techniques. The N-gram technique is used to extract the local context of the news articles, while LSA is used to capture the global context. The temporal features of the news articles are extracted using a sliding window approach, which takes into account the publishing time of the news articles.

The extracted features are then used to train a maximum entropy model, which is used to classify the news articles into different events. The maximum entropy model is chosen due to its ability to handle high-dimensional and sparse data, as well as its capability to model complex nonlinear relationships between the features.

Performance- The performance of the proposed method is evaluated using several evaluation metrics such as precision, recall, and F1-score. The results show that the proposed method outperforms several state-of-the-art methods in terms of event detection and clustering accuracy.

Dataset- The dataset used is a collection of news articles from Chinese news agencies. The dataset was collected over a period of three months and consists of 2,000 news articles with a total of 19,880 sentences.

Year- 2015

C. Paper 3

Author- Pengda Qin, Weiran Xu, William Yang Wang, and William W. Cohen

Title- Unsupervised Event Extraction and Representation Learning from News Articles

Methodology- The authors proposed an unsupervised event extraction and representation learning method, called Unsupervised Event Embedding (UEE). Their approach aimed to capture event-specific information and generate event representations for downstream tasks.

Performance- The results showed that the UEE method outperformed the two baseline methods on both datasets in terms of event extraction and representation learning. The authors also conducted a qualitative analysis of the extracted events to evaluate their quality.

Dataset- The UEE method was evaluated on two different datasets: the English Gigaword dataset and the Chinese Gigaword dataset. The English Gigaword dataset consists of over 9 million news articles, and the Chinese Gigaword dataset consists of over 10 million news articles.

Year- 2018

D. Paper 4

Author- Gaurav Singh and Ritesh Kumar

Title- A Deep Neural Network for Event Extraction from News Articles

Methodology- The proposed deep neural network model for event extraction from news articles in this paper consists of two main components: a convolutional neural network (CNN) and a recurrent neural network (RNN). The CNN is used to extract event trigger words from the input sentences, while the RNN is used to extract the event arguments. The model is trained end-to-end using a cross-entropy loss function.

Performance- The proposed model is evaluated on two benchmark datasets: ACE 2005 and TAC-KBP 2017. The model achieves state-of-the-art performance on both datasets. On the ACE 2005 dataset, the model achieves an F1 score of 54.6%, which is a significant improvement over the previous best result of 50.6%. On the TAC-KBP 2017 dataset, the model achieves an F1 score of 49.4%, which is also a significant improvement over the previous best result of 45.7%.

Dataset- The authors evaluate their proposed model on two benchmark datasets: ACE 2005 and TAC-KBP 2017. The ACE 2005 dataset consists of 599 news articles annotated with event mentions, entity mentions, and temporal expressions. The TAC-KBP 2017 dataset is a newswire dataset that consists of 103 documents annotated with event mentions, entity mentions, and relations between entities. Both datasets are widely used for evaluating event extraction systems. The authors split the datasets into training, validation, and test sets in a standard way for evaluation purposes.

Year- 2018

E. Paper 5

Author- Lei Sha, Lili Mou, Tianyu Liu, Pascal Poupart, Sujian Li, Baobao Chang

Title- Event Extraction via Multi-Level Attention CNNs

Methodology- The proposed model in this paper is a multi-level attention convolutional neural network (CNN) for event extraction from news articles. The model consists of three levels of attention mechanisms: token-level attention, trigger-level attention, and argument-level attention. The token-level attention is used to weight the importance of each word in the input sentence, while the trigger-level attention is used to focus on the most relevant trigger word for the event. The argument-level attention is used to select the most important arguments for the event. The model is trained using a multi-task learning approach, where the trigger and argument extraction tasks are trained jointly.

Performance- The proposed model is evaluated on two benchmark datasets: ACE 2005 and ERE (Event Nugget Detection and Coreference Resolution). The model achieves state-of-the-art performance on both datasets. On the ACE 2005 dataset, the model achieves an F1 score of 53.14%, which

is a significant improvement over the previous best result of 51.21%. On the ERE dataset, the model achieves an F1 score of 46.51%, which is also a significant improvement over the previous best result of 44.84%.

Dataset- The authors evaluate their proposed model on two benchmark datasets: ACE 2005 and ERE. The ACE 2005 dataset is a widely used benchmark dataset for event extraction, consisting of 599 news articles annotated with event mentions, entity mentions, and temporal expressions. The ERE dataset is a more recent benchmark dataset, consisting of a subset of the TAC KBP 2015 and 2016 datasets. The ERE dataset focuses on event nugget detection and coreference resolution. The authors use the standard train-test splits for both datasets and report the performance of their model on the test sets.

Year- 2018

III. Methodology

The methodology used in the above code aims to extract events from news articles and cluster them together based on their semantic similarity. The code uses various natural language processing techniques and machine learning algorithms to achieve this goal.

The first step in the methodology is to extract news articles using the NewsAPI. The API provides access to various news sources, and the code filters the articles based on a user-defined query. The query can be any word or phrase that the user wants to extract news for. The code then downloads the news articles in batches and stores them in a Pandas dataframe.

Next, the code performs event extraction on the news articles. Event extraction is the process of identifying and extracting information about real-world events from unstructured text data. In this code, event extraction is done by using the spaCy library, which is a powerful open-source library for natural language processing. The library provides various pre-trained models for text processing, and the code uses the en core web sm model, which is a small English language model that can perform various NLP tasks such as part-of-speech tagging, named entity recognition, and sentence segmentation.

The event extraction process involves tokenizing the news articles, converting them to lowercase, removing stop words, and lemmatizing the remaining words. After this pre-processing step, the code uses spaCy to extract all the verbs and their corresponding noun phrases from the news articles. These verb-noun phrases represent the events in the articles. For example, in the sentence "The CEO announced a new product," the verb-noun phrase would be "announced a new product." These event phrases are then stored in a list for further processing.

The next step in the methodology is to represent the event phrases as vectors. Vector representation is essential in machine learning algorithms, as it allows for efficient computation and comparison of similarities between data points. In this code, the event phrases are represented as vectors using spaCy's built-in vectorizer, which generates a dense vector representation of the phrases based on their word embeddings. These vectors are then stored in a list for further processing.

After obtaining the event vectors, the code uses a clustering algorithm to group similar events together. In this code, the clustering algorithm used is the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. DBSCAN is a density-based clustering algorithm that groups together data points that are closely packed together while marking outliers as noise. The DBSCAN algorithm requires two parameters to be set: the epsilon value (eps) and the minimum number of points (minsamples). The eps value determines the minimum distance between two points for them to be considered part of the same cluster, while the min samples value determines the minimum number of points required to form a cluster.

To determine the optimal values for the eps and min samples parameters, the code performs a grid search over a range of values for eps and counts the number of clusters formed for each eps value. The optimal eps value is the one that results in the highest number of clusters. The code then runs the DBSCAN algorithm using the optimal eps value and a min samples value of 5, which is a commonly used value in DBSCAN clustering.

After clustering the event vectors, the code assigns a label to each event phrase based on the cluster it belongs to. The labeled event phrases are then stored in a Pandas dataframe along with their corresponding news article.

Finally, the code prints out the number of events in each cluster and displays a few examples of events from a selected cluster.

In conclusion, the methodology used in the above code combines various natural language processing techniques and machine learning algorithms to extract events from news articles and cluster them together based on their semantic similarity. The code uses the NewsAPI to obtain news articles, spaCy to extract event phrases and represent them as vectors, and the DBSCAN to cluster the vectors.

IV. Result and Analysis

TABLE I
TABLE SHOWING CLUSTERS FOR DIFFERENT VALUES OF EPS AND MIN_SAMPLES

eps	**min_ samples**	**Clusters**				
		No. of clusters	***-1***	***0***	***1***	***2***
0.029	3	3	179	1946	3	
0.0165	4	4	1377	743	4	4
0.015	5	4	1712	399	12	5

The results of the event extraction and clustering are summarized as follows:

- The DBSCAN algorithm was applied to the event vectors with an epsilon value of 0.015 and a minimum number of samples of 5. Also I experimented with different values of epsilon and minimum number of samples. These have been shown in the above table.

- In case of eps=0.015 and minimum number of samples=5, I got 4 clusters.
- Here cluster number -1 represents noisy data.
- The descriptions that were clustered together share similar topics and content. Some of the topics include sports, politics, health, and technology.
- The algorithm was able to group together similar news articles and identify topics and themes in the dataset.

V. Future Scope

This approach appears to be effective in extracting events and clustering news articles based on their content. However, the choice of epsilon value and minimum number of samples can have a significant impact on the results, and these parameters need to be carefully tuned for optimal performance. Additionally, the quality of the results is highly dependent on the quality of the text preprocessing and feature extraction techniques used, and there is always room for improvement in these areas.

VI. Code

Code is available at: Github **https://github.com/Siddhant-Agarwal4583/NLP_Project**

References

[1] Felix Hamborg, Corinna Breitinger and Bela Gipp, Giveme5W1H: A Universal System for Extracting Main Events from News Articles, 2019.

[2] Jin Zhang, Tao Li and Wei Xu, Event Detection and Clustering Using Maximum Entropy Based on Semantic and Temporal Features, 2015.

[3] Pengda Qin, Weiran Xu, William Yang Wang, and William W. Cohen, Unsupervised Event Extraction and Representation Learning from News Articles, 2018.

[4] https://paperswithcode.com/paper/giveme5w1h-a-universal-system-for-extracting

[5] https://github.com/fhamborg/Giveme5W1H

[6] https://github.com/daniel-aracquine/event-extraction-nlp