5W1H Extraction: Cause & Method Extractor

Iulia-Diana Groza

April 23, 2024

Abstract

This theoretical report analyses the extraction of "Why?" (Cause) and "How?" (Method) components, as part of 5W1H (Who?, What?, When?, Where?, Why?, and How?), in text using Natural Language Processing (NLP). Traditional NLP methods, often limited by linguistic rules and shallow machine learning models, have evolved with the integration of deep learning techniques, offering significant improvements in precision and adaptability. We particularly examine the Giveme5W1H tool, a state-of-the-art system developed by Hamborg et al., which utilises a series of analysers equipped with heuristic and machine learning techniques to automate the extraction of 5W1H elements from English news articles, serving in news summarization and event reconstruction. The report synthesizes and compares current methodologies, ending with an evaluation of the effectiveness of the Giveme5W1H tool.

1 Introduction

Natural Language Processing (NLP) continuously evolves to address complex information extraction challenges within large text corpora. A particularly demanding aspect of NLP is the extraction of the *5W1H elements —Who?, What?, When?, Where?, Why?, and How?* — which fundamentally describe the main components of textual information. The extraction of "Why?" and "How?" is critical as these elements usually contain causal and procedural insights that are essential for real-life applications such as news summarization, event reconstruction, psychotherapy, journalism, and decision support systems.

Traditional methods of NLP rely heavily on linguistic rules and shallow machine learning models, which often fall short when faced with the ambiguity and variability of natural language. However, deep learning has provided significant improvements in this domain. For instance, Nurdin and Maulidevi (2018) demonstrated the effectiveness of combining *Convolutional Neural Networks* (CNNs) with *Bidirectional Long Short-Term Memory* (LSTM) networks for 5W1H extraction from Indonesian news articles, achieving notable precision through deep semantic feature extraction [1].

The *Giveme5W1H tool*, introduced by Hamborg et al., stands out as a state-of-the-art system designed to automate the extraction of 5W1H elements from English-language news texts [2]. It applies a series of analyzers, each focusing on one of the 5W1H questions, driven

by a combination of heuristic and machine learning techniques to address the challenge of extracting structured information from unstructured text sources.

This theoretical report aims to present and explore the methodology of extracting the cause ("Why?") and the method ("How?") chains, by reviewing the existing literature. Additionally, we will analyse the use of the Giveme5W1H tool developed by Hamborg et al., focusing on how their open-source system [3] can be leveraged for extracting the cause and the method from the main event of a news article. These two components hold the key to understanding the motives and processes behind events, which are essential for deeper analysis and applications requiring detailed understanding and response planning. Jang and Woo's research on unified user-centric context representation in ubiquitous computing environments highlights the importance of structuring complex information according to the 5W1H framework, as it aids in the semantic interpretation and automation of context-aware services [4].

2 Literature Review

In their seminal work, *"Giveme5W1H: A Universal System for Extracting Main Events from News Articles"* (2018), Hamborg et al. developed the Giveme5W1H tool, which utilises *four* distinct analysers for all 5W1H components, using a mix of heuristic and machine learning techniques. While comprehensive, its efficacy in processing the nuanced "Why?" and "How?" components across varied datasets could benefit from deeper analysis [2].

Jang and Woo, in *"5W1H: Unified user-centric context."* (2005), focus on classifying user-centric context within a 5W1H framework to enhance applications in ubiquitous computing. Their model's relevance to direct text extraction, particularly for implicit elements like "Why?" and "How?", remains limited, suggesting potential for further application enhancements [4].

Nurdin and Maulidevi's approach, from *"5W1H Information Extraction with CNN - Bidirectional LSTM"* (2018), uses CNNs and Bidirectional LSTMs to refine 5W1H extraction, excelling particularly with "Why?" and "How?" due to its robust handling of contextual and temporal dependencies [1]. This model's capacity for deep semantic learning offers significant improvements over more static or heuristic-driven approaches.

Comparatively, Nurdin and Maulidevi's model provides a deeper semantic grasp than the Giveme5W1H tool, providing customization options without manual intervention. The structured context interpretation proposed by Jang and Woo could further enrich these models, particularly in nuanced information extraction tasks.

It is important to note that Jang and Woo perform the 5W1H extraction regardless of the domain of the input text, creating one more advantage for their work, by offering a general purpose for their tool. Both Giveme5W1H and the CNN - Bidirectional LSTM approach are destined for use on news article, so the results are comparable. However, there is a major difference between the two works: the language of the source text. Giveme5W1H potentially benefits more from leveraging NLP techniques, as existing tools are more powerful on English resources. Consequently, deep learning solutions compensate in this sense for Nurdin and Maulidevi's work on Indonesian news articles.

Harnessing neural attention mechanisms and transfer learning techniques suggest further

enhancements. For example, the Transformer model introduced by Vaswani et al., in *"Attention is All You Need"* (2017), demonstrates how focusing on relevant text portions can improve extraction accuracy [5], while BERT's approach to pre-trained models, as discussed by Devlin et al. in *"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"* (2018), shows promise in understanding complex narrative structures across varied texts [6].

3 Methodology

This section elaborates on the methodology of the Giveme5W1H tool developed by Hamborg et al., with a particular emphasis on the extraction of the *"Cause"* and *"Method"* ("Why?" and "How?") chains, for understanding the underlying dynamics and procedures of news events [2].

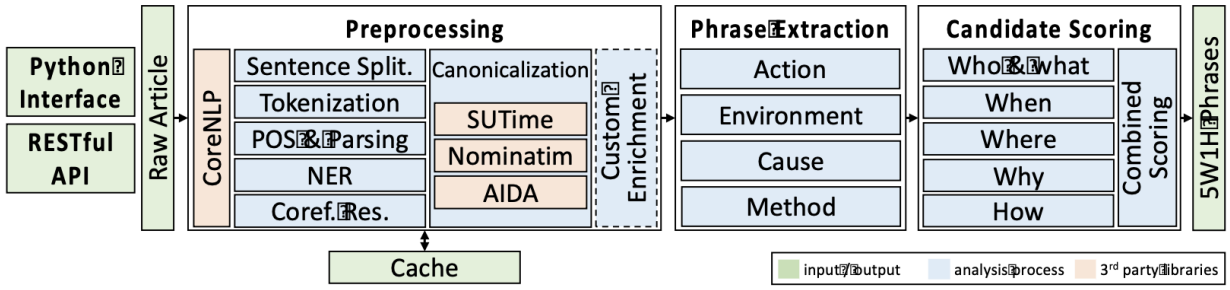

Figure 1: The three-phases analysis pipeline preprocesses a news text, finds candidate phrases for each of the 5W1H questions, and scores these. Source: *"Giveme5W1H: A Universal System for Extracting Main Events from News Articles"* [2]

Giveme5W1H initiates with a comprehensive preprocessing of the news article, utilizing *Stanford CoreNLP* to perform *sentence splitting*, *tokenization*, *POS-tagging*, and *Named Entity Recognition* (NER).

The core of Giveme5W1H's functionality lies in its modular design, where separate extractors are dedicated to each of the 5W1H questions. This modular approach allows for focused improvements and adaptations to each component without affecting the overall system. There are four chains for the main event extraction:

- *Action:* Who?, What?
- *Environment:* When?, Where?
- *Cause:* Why?
- *Method:* How?

The *"Cause"* module searches for linguistic patterns that indicate causality, such as causal conjunctions (e.g., "because", "due to") and verbs (e.g., "cause", "result in"). By analyzing the syntactic structures within sentences, it identifies and extracts phrases that provide explanations or reasons behind the events described in the news article.

The *"Method"* module is focused on detailing the process or manner in which events occur. It looks for adverbial phrases and specific verb constructions that describe actions. The complexity of extracting the "How?" component stems from the diverse ways actions can be described, requiring the tool to interpret a wide range of linguistic cues.

After phrase extraction, the authors integrate a *scoring system* that evaluates and ranks the extracted phrases based on their relevance, accuracy and linguistic context in answering the respective 5W1H questions. This scoring system is developed based on multiple heuristics, such as the frequency of phrase occurrence, its position within the article, and its contextual relevance. The scoring system is detailed in Tables 1 and 2, which list the scoring factors and their respective weights for the "Why?" and "How?" questions.

The scoring of "Why?" answers involves evaluating the clarity and directness of causation indicated by the extracted phrases. Key factors include the syntactic position of the causative keyword and the completeness of the causal clause.

i	$w_{\text{why},i}$	$s_{\text{why},i}$
0 (position)	.56	$\text{pos}(c)$
1 (type)	.44	$\text{CT}(c)$

Table 1: Weights and scoring factors for 'why' phrases. Source: *"Giveme5W1H: A Universal System for Extracting Main Events from News Articles"* [2]

Scoring for the "How?" answers assesses how comprehensively the methods or procedures are described in the extracted phrases. Important factors include the detail provided by the adverbial modifiers and the contextual relevance of the procedural description.

i	$w_{\text{how},i}$	$s_{\text{how},i}$
0 (position)	.23	$\text{pos}(c)$
1 (frequency)	.14	$f(c)$
2 (type)	.63	$\text{TM}(c)$

Table 2: Weights and scoring factors for 'how' phrases. Source: *"Giveme5W1H: A Universal System for Extracting Main Events from News Articles"* [2]

Finally, Giveme5W1H incorporates parameter learning to refine its extraction capabilities. Parameter learning is a critical aspect that enables the system to adjust and optimize the weights used in its scoring system based on feedback and performance evaluation. This is particularly relevant for the "Why?" and "How?" chains, where the subtlety of causal and procedural information often requires a nuanced approach to extraction. The learning algorithm iteratively adjusts the parameters, such as the weights for position and type within the scoring tables, to minimize extraction errors and improve precision. By leveraging annotated datasets and performance metrics from previous extractions, the system can learn which patterns are most indicative of accurate information for each of the 5W1H elements.

4 Experiments and Results

The evaluation used a dataset of 120 articles, with three assessors rating the relevance of the extracted 5W1H phrases on a three-point scale. The results showed an overall mean average generalized precision (MAgP) of 0.73 across all categories, with a higher precision of 0.82 when considering only the first four W questions (Who, What, When, Where), which are often deemed sufficient to summarize the main event.

The performance of the system varied across different categories of news, with politics articles showing the highest extraction quality. Sports articles also performed well, despite the complexity due to the multiple events often described within them. Entertainment and tech articles posed challenges, mainly due to their background information and the indirect way in which they often report events.

Question	**ICR**	**Bus**	**Ent**	**Pol**	**Spo**	**Tec**	**Avg.**
Who	.93	.98	.88	.89	.97	.90	.92
What	.88	.85	.69	.89	.84	.66	.79
When	.89	.55	.91	.79	.81	.82	.78
Where	.95	.82	.63	.85	.79	.80	.78
Why	.96	.48	.62	.42	.45	.42	.48
How	.87	.63	.58	.68	.51	.65	.61
Avg. all	.91	.72	.72	.75	.73	.71	.73
Avg. 4W	.91	.80	.78	.86	.85	.80	.82

Table 3: ICR and MAgP-Performance of Giveme5W1H. Source: *"Giveme5W1H: A Universal System for Extracting Main Events from News Articles"* [2]

The assessment revealed that the 'Why' and 'How' components were particularly challenging to extract accurately. This is attributed to the articles often only implying causes and methods rather than stating them explicitly. These findings align with the known difficulty in NLP of extracting implied information, which requires sophisticated semantic interpretation that current technologies may not fully support.

To improve the extraction of the 'What' component, there's a plan to develop separate extraction methods for 'Who' and 'What' to ensure the top candidates for both fit together semantically. For temporal ('When') and locational ('Where') information, the researchers intend to use linked data from knowledge graphs like YAGO when explicit mentions are absent from the article.

Improvements in the precision of 'Why' and 'How' extractions are also planned by setting score thresholds to avoid false positives and employing more advanced NLP methods for better semantic interpretation.

5 Conclusions

This theoretical report has presented an in-depth examination of the Giveme5W1H tool for the extraction of the "Cause" and "Method" components from news articles. Our analysis highlights the tool's use of machine learning and heuristic techniques to address the task

of 5W1H extraction. While the tool shows commendable performance across various news domains, with a notable mean average generalized precision (MAgP), it particularly excels in the extraction of "Who?", "What?", "When?", and "Where?" components.

The results reinforce the notion that the extraction of "Why?" and "How?" remains a significant challenge within the field of NLP, reflecting the nuanced nature of causation and methodology in textual data. Despite the high intercoder reliability (ICR) indicating consistency in the tool's performance, the relatively lower precision in extracting "Why?" and "How?" suggests that there is still considerable room for improvement.

For future development, enhancing the tool's ability to infer implicit information stands out as a primary objective. The integration of emerging technologies, such as neural attention mechanisms and contextually-aware algorithms, could provide the means to capture the subtleties of causal and procedural language. Additionally, expanding the tool to process languages beyond English could greatly increase its applicability and usefulness in global contexts.

Furthermore, the open-source nature of the Giveme5W1H tool encourages collaborative improvements and adaptations, providing a solid foundation for the broader research community to contribute to the evolution of this technology. As the tool matures and incorporates more sophisticated NLP techniques, its utility in real-world applications is likely to expand, offering more precise and contextually relevant extractions that can benefit journalists, researchers, and analysts alike.

In conclusion, the work of Hamborg et al. represents a significant step forward in automated 5W1H extraction. However, the journey towards perfecting this tool and the broader challenges of NLP is an ongoing process, demanding continuous innovation and research. It is through such endeavors that we can aspire to create systems capable of understanding the complexities of human language, providing valuable insights into the vast corpus of textual information that surrounds us.

References

[1] Nurdin, A., & Maulidevi, N. U. (2018). *5W1H Information Extraction with CNN-Bidirectional LSTM.* Journal of Physics: Conference Series, 978, 012078. IOP Publishing. doi:10.1088/1742-6596/978/1/012078

[2] Hamborg, F., Lachnit, S., Schubotz, M., Heinz, T., Gipp, B., & Aizawa, A. (2018). *Giveme5W1H: A Universal System for Extracting Main Events from News Articles.* Proceedings of the 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), [pages]. IEEE. doi:10.1109/JCDL.2019.00100

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

[4] Jang, Seiie, and Woontack Woo. *5W1H: Unified user-centric context.* The 7th International Conference on Ubiquitous Computing. 2005.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is All You Need.* Advances in Neural Information Processing Systems, 30, [pages]. doi:10.5555/3295222.3295349

[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* arXiv preprint arXiv:1810.04805.

Acknowledgement: This work is the result of my own activity, and I confirm I have neither given, nor received unauthorized assistance for this work. I declare that I did not use generative AI or automated tools in the creation of content or drafting of this document.