5W1H Extraction: Cause & Method Extractor

Iulia-Diana Groza

5W1H

Who? What? When? Where? **Why?** How?

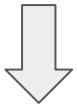
Taliban attacks German consulate in northern Afghan city of Mazar-i-Sharif with truck bomb

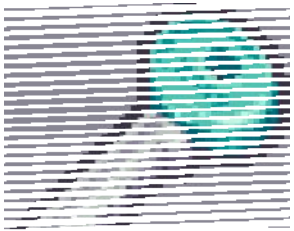
The death toll from a powerful Taliban truck bombing at the German consulate in Afghanistan's Mazar-i-Sharif city rose to at least six Friday, with more than 100 others wounded in a major militant assault.

The Taliban said the bombing late Thursday, which tore a massive crater in the road and overturned cars, was a "**revenge attack**" for US air strikes this month in the volatile province of Kunduz that left 32 civilians dead. […]

5W1H in NLP

Extract semantic roles from text, mapping words or phrases to attributes in the 5W1H, aiding in information structuring and analysis.





Enhance comprehension of text contexts and relationships, and data extraction in various applications (e.g. journalism, social media analysis).

Challenges in 5W1H Extraction

- **Complexity of "Why?" and "How?":** Traditional NLP methods struggle with the implicit nature of causal and procedural information.
- **Limitations of Linguistic Rules:** Rule-based systems fail to capture the nuanced context of "Why?" and "How?".
- **Deep Learning Advantages:** Transformers, CNNs and LSTMs offer improved semantic understanding, with higher precision.

Literature Review

Giveme5W1H: A Universal System for Extracting Main Events from News Articles (2018), Hamborg et al.

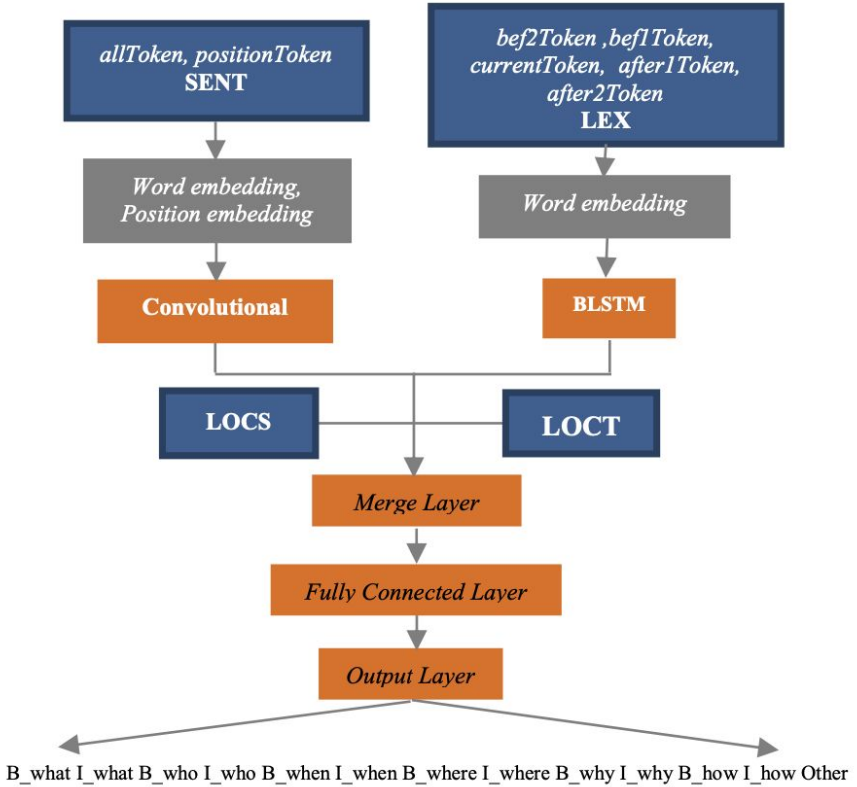
Research Contribution:

- Extracted **"Why?"** and **"How?"** by recognizing implicit causal and procedural information throughout **English news articles**.
- Introduced a **scoring system** that evaluates the relevance and accuracy of phrases, aiming to reduce false positives and refine the system's performance specifically for "Why?" and "How?" questions.

5W1H Information Extraction with CNN-Bidirectional LSTM (2018), Nurdin & Maulidevi

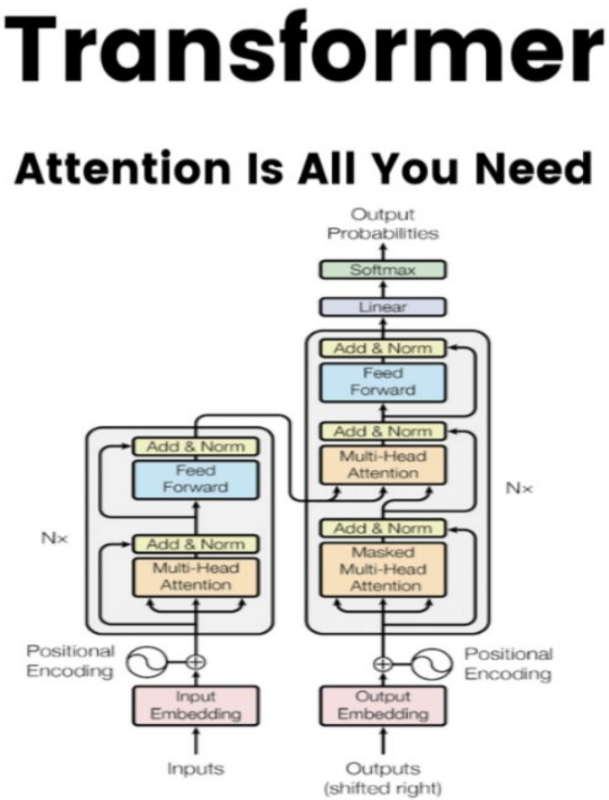
Research Contribution:

- Demonstrated effectiveness of **deep learning (CNN and LSTM)** models in extracting **"Why?"** and **"How?"** from **Indonesian news articles**.
- Highlighted the need for semantic feature extraction to understand the underlying context.



Attention is All You Need (2017), Vaswani et al.

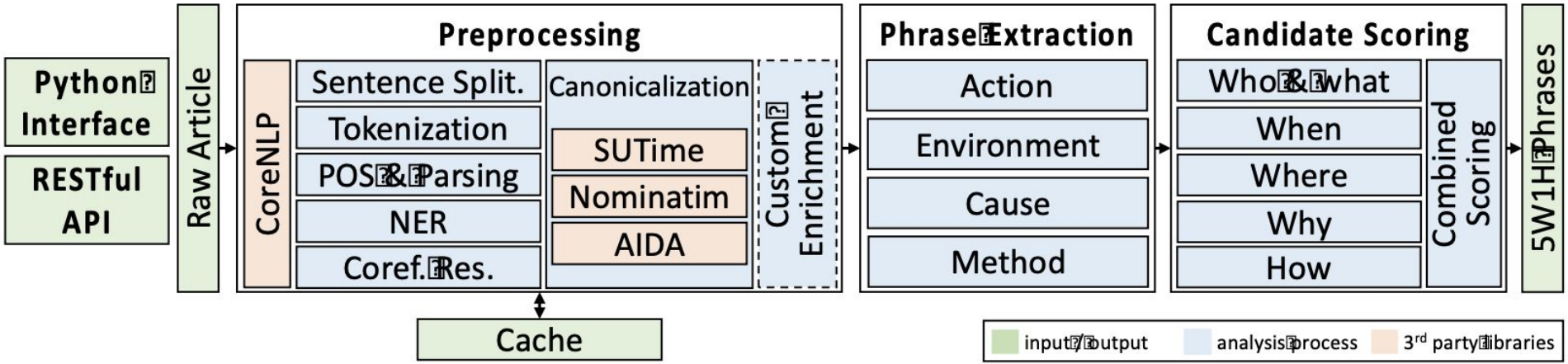
- **Innovative Mechanism:** Introduces the **Transformer** model - self-attention mechanisms to weigh the importance of different words in a sentence **without relying on recurrent layers**.
- **Improved Contextual Awareness:** Captures **dependencies** across all words in the text **simultaneously** in the context of **"Why?"** and **"How?"**.
- **Enhanced Precision:** Focus on relevant parts of text **dynamically** by understanding **underlying patterns** and connections in complex narrative structures.



Giveme5W1H

https://github.com/fhamborg/Giveme5W1H

Methodology



Scoring System

- **Key Components of the Scoring System:**
  - **Weighted Factors:** Each candidate phrase is scored based on the semantic relevance, proximity to keywords, and contextual alignment with the query.
  - **Dynamic Adjustment:** Weights for different factors are dynamically adjusted based on the learning from previous extractions to continuously improve accuracy.
- **Impact on 'Why?' and 'How?' Extractions:**
  - **“Why?” Scoring:** Causal relationships; weights are heavily placed on semantic indicators of causation such as **resultative verbs** and **causal conjunctions**.
  - **“How?” Scoring:** Procedural clarity and detail; higher weights assigned to phrases that **clearly describe** the method or process of an event.

Scoring System

Table 4: Weights and scoring factors for 'why' phrases

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

Table 5: Weights and scoring factors for 'how' phrases

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

Parameter Learning

- **Fine-tune the weights:** Adjust parameters based on **feedback** from the tool's performance in **real extraction tasks**.
- **Precision Improvement:** By iteratively optimizing parameters, the tool learns to better distinguish relevant from irrelevant information, particularly improving the extraction of complex **"Why?"** and **"How?"** questions.
- **Reduced Errors:** **Continuous learning** helps in **minimizing false positives** and enhancing the overall reliability of the extraction outcomes, ensuring that the extracted information is both accurate and contextually appropriate.

Experiments and Results

Table 6: ICR and MAgP-Performance of Giveme5W1H

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**

Conclusions

- **Complexity of "Why?" and "How?":** Traditional NLP methods struggle with the implicit nature of causal and procedural information.
- **No existing tools for Romanian Language:** Most solutions are developed in English. Developing a Romanian tool requires great linguistic resources.
- **Importance of open sourcing resources:** Giveme5W1H is the only 5W1H extraction tool from the literature that is open sourced.

Demo

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). IEEE. doi:10.1109/JCDL.2019.00100
3. 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. doi:10.5555/3295222.3295349
4. https://github.com/fhamborg/Giveme5W1H
5. https://navigossearch.com/what-is-5w1h-the-meaning-and-application-of-the-5w1h-method-in-various-fields