Erschienen in: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries - JCDL '18 / Chen, Jiangping et al. (Hrsg.). - New York : ACM Press, 2018. - S. 339-340. - ISBN 978-1-4503-5178-2
https://dx.doi.org/10.1145/3197026.3203899

Extraction of Main Event Descriptors from News Articles by Answering the Journalistic Five W and One H Questions

Felix Hamborg, Corinna Breitinger, Moritz Schubotz, Soeren Lachnit, Bela Gipp
Department of Computer and Information Science
University of Konstanz, Germany
{firstname.lastname}@uni-konstanz.de

ABSTRACT

The identification and extraction of the events that news articles report on is a commonly performed task in the analysis workflow of various projects that analyze news articles. However, due to the lack of universally usable and publicly available methods for news articles, many researchers must redundantly implement methods for event extraction to be used within their projects. Answers to the journalistic five W and one H questions (5W1H) describe the main event of a news story, i.e., who did what, when, where, why, and how. We propose Giveme5W1H, an open-source system that uses syntactic and domain-specific rules to extract phrases answering the 5W1H. In our evaluation, we find that the extraction precision of 5W1H phrases is $p = 0.64$, and $p = 0.79$ for the first four W questions, which discretely describe an event.

KEYWORDS

News Event Detection, 5W1H Extraction, 5W1H Question Answering, Reporter's Questions, Journalist's Questions, 5W QA.

1 INTRODUCTION AND RELATED WORK

Extraction of a news article's main event is a fundamental analysis task required for a broad spectrum of use cases. For instance, news aggregators must identify the main event to cluster related articles [2, 3], e.g., articles reporting on the same event. Other disciplines also analyze the events reported on in articles, e.g., in so called frame analyses, social scientists identify how the media is reporting on certain events. However, no method is publicly available that extracts explicit descriptors of the main event. We define *explicit event descriptors* as the properties that occur in a text describing an event, e.g., the text phrases in a news article that enable a news reader to understand what the article is reporting on.

The clear majority of current approaches suffer from at least one of three shortcomings. First, they detect events only implicitly, e.g., by employing topic modeling, but do not extract phrases or properties that explicitly describe the article's main event [3]. The second category of approaches does not extract universally usable descriptors, but is specialized on the extraction of task-specific event properties, such as the number of protestors in a demonstration [4]. Approaches of the third category extract explicit event descriptors but are not publicly available [5].

Journalists typically answer the journalistic five W and one H questions (5W1H), i.e., *who* did *what*, *when*, *where*, *why*, and *how*, within the first few sentences of an article to inform the readers of the main event. For instance, the headline of a news article reporting on a terrorist attack in Afghanistan answers four of the 5W1H questions: "Taliban attacks German consulate in northern Afghan city of Mazar-i-Sharif with truck bomb" The highlighted phrases answer the questions who did what, where, and how; 'when' and 'why' are answered in the remainder of the article. Due to their descriptiveness of an article's main event, we focus our research on the extraction of the journalistic 5W1Hs.

2 EXTRACTION OF 5W1H PHRASES

Giveme5W1H is an open-source main event retrieval system for news articles. The system uses syntactic and domain-specific rules to extract the 5W1H phrases in a three-phase analysis pipeline depicted in Figure 1. The system builds on Giveme5W [1], and improves the extraction performance by addressing multiple of the future work directions: Giveme5W1H uses coreference resolution, question-specific semantic distance measures, combined scoring of candidates, and extracts phrases for the 'how' question.

In the first phase, *preprocessing*, Giveme5W1H performs state-of-the-art NLP, and canonicalization to bring all named entities (NE) in their normalized form. During canonicalization we parse dates written in natural language into canonical dates (TIMEX3),

This work has been supported by the Carl Zeiss Foundation.

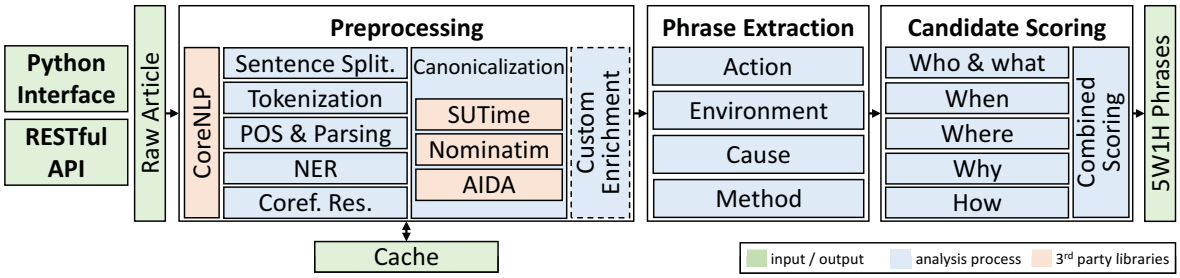

Figure 1: The three-phase analysis pipeline (1) preprocesses a news text, (2) finds candidate phrases for each of the 5W1H questions, and (3) scores these. Giveme5W1H can easily be accessed in Python and via a RESTful API.

perform geocoding, and perform NE recognition and disambiguation (NERD) to link all NEs to concepts in the YAGO graph.

The second phase, *phrase extraction*, uses four extraction chains to retrieve candidate phrases: (1) the *action* chain extracts phrases for the journalistic 'who' and 'what' questions, (2) *environment* extracts 'when' and 'where', (3) *cause* extracts 'why', and (4) *method* extracts 'how'. We extract as 'who' candidates all subjects, i.e., the first noun phrase (NP) from each sentence, and as 'what' candidates their respective predicates. To determine the 'when' candidates, we take the TIMEX3 instances, which Giveme5W1H extracts during preprocessing. Similarly, we take the geocodes as 'where' candidates. To find 'why' candidates, we look for three types of cause-effect indicators: causal conjunctions, causative adverbs, and causative verbs. We then use a set of syntax rules to find the respective causal phrase. To find 'how' candidates, we analyze copulative conjunctions, adjectives and adverbs. Often, sentences with a copulative conjunction, such as "after [the train came off the tracks]", contain a method phrase in the clause that follows the copulative conjunction. To improve extraction quality, we post-process all candidates for each question, e.g., by truncating long phrases or discarding too short phrases.

The third phase, *candidate scoring*, aims to determine the best candidate of each 5W1H question. First, we score candidates independently for each of the 5W1H questions. Therefore, we devise domain-specific scoring rules. For instance, to score 'who' candidates, we define three scoring factors, which we motivate from journalistic writing concepts, such as the inverse pyramid and journalistic hooks: the candidate shall occur in the article (1) *early* and (2) *often*, and (3) contain an *NE*. If a candidate contains an NE, we will score it higher, since in news articles, the actors involved in events are often NEs, e.g., politicians. Lastly, we perform a combined scoring that adjusts the score of a given candidate, depending on the properties of top candidates of other questions, e.g., the method by which an action was performed ('how') is usually described in the same or an adjacent sentence as the action ('what').

Giveme5W1H returns the top candidate phrase for each question, including the normalized data from canonicalization.

3 RESULTS

We use the evaluation dataset by Hamborg et al. [1], which consists of 60 articles in the categories business (*Bus*), entertainment (*Ent*), politics (*Pol*), sport (*Spo*), and tech (*Tec*). We asked three assessors to judge the relevance of each answer on a 3-point scale (non-relevant, partially relevant, and relevant). Table 1 shows the mean average generalized precision (MAgP). The MAgP over all categories and questions was 0.64. If only considering the first 4Ws, which are sufficient to uniquely represent an event, the overall MAgP was 0.79. Extracting the answers to 'why' and 'how' performed worse, since news articles often only imply causes and methods. The extraction performance is similar to state-of-the-art approaches [5], but a direct comparison is not possible due to the non-availability of methods and datasets (see Section 1).

Table 1: MAgP-Performance of Giveme5W1H

Question	Bus	Ent	Pol	Spo	Tec	Avg.
Who	.98	.88	.85	.97	.86	.91
What	.77	.67	.89	.83	.63	.75
When	.55	.91	.79	.77	.82	.77
Where	.82	.63	.85	.77	.68	.75
Why	.36	.18	.32	.33	.40	.32
How	.25	.36	.45	.27	.46	.36
Avg. all	.62	.61	.69	.66	.64	**.64**
Avg. 4W	.78	.65	.84	.83	.75	**.79**

4 CONCLUSION

We proposed *Giveme5W1H*, the first open-source system that extracts answers to the journalistic 5W1H questions, i.e., *who* did *what, when, where, why*, and *how*, to describe a news article's main event. Giveme5W1H achieved a mean average generalized precision (MAgP) of 0.64 for all questions, and an MAgP of 0.79 in answering the questions who, what, when, and where, which can uniquely represent an event. The code of Giveme5W1H and the evaluation dataset are available under an Apache 2 license on GitHub: **https://github.com/fhamborg/Giveme5W1H**

REFERENCES

[1] Hamborg, F. et al. 2018. Giveme5W: Main Event Retrieval from News Articles by Extraction of the Five Journalistic W Questions. *Proceedings of the iConference 2018* (Sheffield, UK, 2018).

[2] Hamborg, F. et al. 2017. Identification and Analysis of Media Bias in News Articles. *Proceedings of the 15th International Symposium of Information Science* (2017).

[3] Hamborg, F. et al. 2017. Matrix-based News Aggregation: Exploring Different News Perspectives. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*. 10, 17 (2017).

[4] Oliver, P.E. and Maney, G.M. 2000. Political Processes and Local Newspaper Coverage of Protest Events: From Selection Bias to Triadic Interactions. *American Journal of Sociology*. 106, 2 (2000), 463–505.

[5] Parton, K. et al. 2009. Who, what, when, where, why?: comparing multiple approaches to the cross-lingual 5W task. *Proc. of the Joint Conf. of the 47th Annual Meeting of the ACL and the 4th Int. Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1* (2009), 423–431.