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

Motivation & Background

- **Event extraction** is required in many projects analyzing news: news aggregation, clustering related articles, summarization, or manual frame analyses in the social sciences
- Shortcomings of state-of-the-art approaches:
  - detect events only **implicitly** [1]
  - extract only **task-specific properties** [1,2]
  - are **not publicly available** [3,4,5,6]
- Disadvantages to the research community
  - **Redundant** implementation **efforts**
  - **Non-optimal accuracy,** since event extraction is necessary but not the final aim

Research Objectives

Devise an automated method to extract the main event of a single news article.

1. **Extract explicit event descriptors**
2. **Exploit characteristics of news articles**
3. **Publicly available**

5W1H Event Descriptors

- Journalistic five W and one H questions (5W1H) describe the main event of an article: **who** did **what**, **where**, **when**, **why**, and **how**?

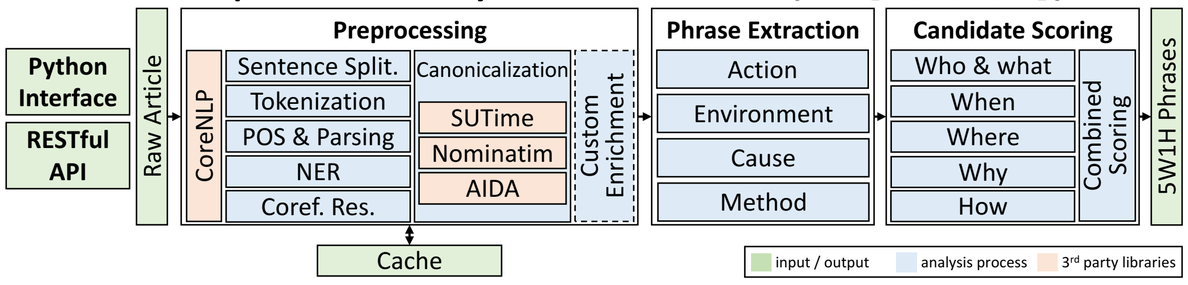
Taliban attacks German consulate in northern Afghan city of Mazar-i-Sharif with truck bomb

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. […]

Approach

- Syntax-based 5W1H extraction using a three-phase analysis workflow (cf. [3,4,5,6])



Phrase Extraction

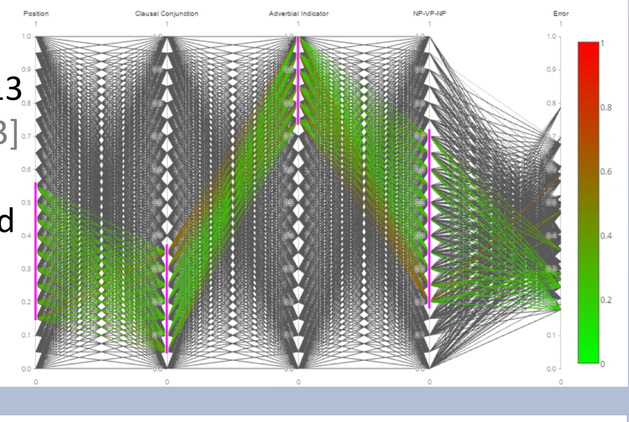
- **Who**: Subject of each sentence (first NP of S)
- **What**: VP to the right of the extracted who-phrase in parse tree
- **Where**: Named entities (NEs) tagged as location
- **When**: TIMEX3 instances extracted by SUTime [8]
- **Why**: POS-patterns (NP-VP-NP) & action verb (e.g., "result of" [10]); tokens (e.g., cause indicating adverbs [9], such as "therefore")
- **How**: Copulative conjunctions ("after [the train came off the tracks]") [11]

Candidate Scoring

- **Who**: (1) **early** in the article ("inverse pyramid" [7], most important info first), (2) **often** (more likely involved in the main event), (3) contain an **NE**
- **What:** same score as adjacent who-phrase due to strong relation
- **Where**: (1) **early**, (2) **often**, (3) **contained** in other locations, (4) **specificity**
- **When:** (1) **early**, (2) **often**, (3) **closeness** to article date, (4) **duration**
- **Why:** (1) **early**, (2) **causal type** (bi-clausal > starts with RB > else)
- **How**: (1) **early**, (2) **often**, (3) **method type** (cop. conj. > else)

Optimal Parameter Configuration

- **Learning dataset**: 3 coders annotated 5W1H in 100 articles from 13 major US and UK news outlets [13]
- ICR = 0.81
- Automatically compared extracted 5W1H with learning dataset
- Found optimum by testing all parameter configurations



Results

- Multi-grade relevance assessment: non-relevant, partially rel., rel.
- Three human assessors rated 60 randomly sampled news articles from the BBC dataset [12]
- Precision (P) = **0.64** & P(4W) = **0.79**
- Comparison with state of the art: P(5W) = 0.65 [3]; MAgP(5W) = 0.89 [6] but dataset was specifically prepared for 5W extraction and not sampled

Property	ICR	Business	Entertainment	Politics	Sports	Tech	Avg.
Who	.92	.98	.88	.85	.97	.86	.91
What	.88	.77	.67	.89	.83	.63	.75
When	.88	.55	.91	.79	.77	.82	.77
Where	.94	.82	.63	.85	.77	.68	.75
Why	.97	.36	.18	.32	.33	.40	.32
How	.87	.25	.36	.45	.27	.46	.36
Avg. all	**.91**	**.62**	**.61**	**.69**	**.66**	**.64**	**.64**
Avg. 4W	**.91**	**.78**	**.65**	**.84**	**.83**	**.75**	**.79**

Project: github.com/fhamborg/Giveme5W1H
Email: felix.hamborg@uni-konstanz.de

References

[1] Tanev, H. et al. 2008. Real-time news event extraction for global crisis monitoring. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2008), 207–218. **[2]** 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. **[3]** Parton, K. et al. 2009. Who, what, when, where, why?: comparing multiple approaches to the cross-lingual 5W task. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1 (2009), 423–431. **[4]** Wang, W. et al. 2010. Chinese news event 5w1h elements extraction using semantic role labeling. Information Processing (ISIP), 2010 Third International Symposium on (2010), 484–489. **[5]** Yaman, S. et al. 2009. Classification-based strategies for combining multiple 5-w question answering systems. INTERSPEECH (2009), 2703–2706. **[6]** Yaman, S. et al. 2009. Combining semantic and syntactic information sources for 5-w question answering. INTERSPEECH (2009), 2707–2710. **[7]** Christian, D. et al. 2014. The Associated Press stylebook and briefing on media law. The Associated Press. **[8]** Chang, A.X. and Manning, C.D. 2012. SUTime: A library for recognizing and normalizing time expressions. LREC. iii (2012), 3735–3740. **[9]** Khoo, C.S.G. et al. 1998. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. Literary and Linguistic Computing. 13, 4 (1998), 177–186. **[10]** Girju, R. 2003. Automatic detection of causal relations for question answering. Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12 (2003), 76–83. **[11]** Oxford English 2009. Oxford English Dictionary. Oxford University Press. **[12]** Greene, D. and Cunningham, P. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. Proceedings of the 23rd international conference on Machine learning (2006), 377–384. **[13]** Hamborg, F. et al. 2017. news-please: A Generic News Crawler and Extractor. Proc. of the 15th International Symposium of Information Science (2017), 218–223.