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

Devising 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, when, where, why, and how?

References


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-causal > 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(SW) = 0.65 [3]; MAGP(SW) = 0.89 [6]

--- | --- | --- | --- | --- | --- | --- | ---
Who | .92 | .98 | .98 | .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 | .81 | .62 | .61 | .69 | .66 | .64 | .64
Avg. 4W | .91 | .78 | .65 | .84 | .83 | .75 | .79

Project: github.com/fhamborg/Giwe5W1H
Email: felix.hamborg@uni-konstanz.de