Giveme5W1H

A Universal System for Extracting Main Events from News Articles

Felix Hamborg[1], Corinna Breitinger[1], Bela Gipp[2]

[1]Dept. of Computer and Information Science
University of Konstanz, Germany

[2]Knowledge and Data Engineering
University of Wuppertal, Germany

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. […] The suicide attacker rammed his explosives-laden car into the wall […].

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. [...] The suicide attacker **rammed his explosives-laden car** into the wall [...].

Who did **what**, **where**, **when**, **why**, and **how**?

Motivation

- News texts
  - answer the five journalistic W and one H questions (5W1H)
  - to quickly inform readers of the main event

- **5W1Hs useful** for various applications
  - Event detection
  - Finding related articles (clustering)
  - Summarization
  - Other sciences, e.g., frame analyses in the social sciences
  - …

Content

- Background
- Methodology
- Evaluation and results
- Conclusion

Background

Event Extraction from News Articles

www.isg.uni.kn

- Current methods
  - extract events **implicitly** (topic modeling, clustering) [2,6,27,32]
  - extract **task-specific** properties [26,32]
  - are **not publicly available** (but extract explicit event descriptors) [29,34-36]
    - Sufficient quality: accuracy ranges from 0.65 [29] to 0.89 [36]

- Disadvantages to the research community
  - **Redundant work** for a common task
  - **Non-optimal accuracy**

Research Objective

www.isg.uni.kn

Devise a method that extracts the main event of a single news article

- *explicit main event descriptors*
- that are *usable* by later tasks in the analysis workflow

- *publicly available*

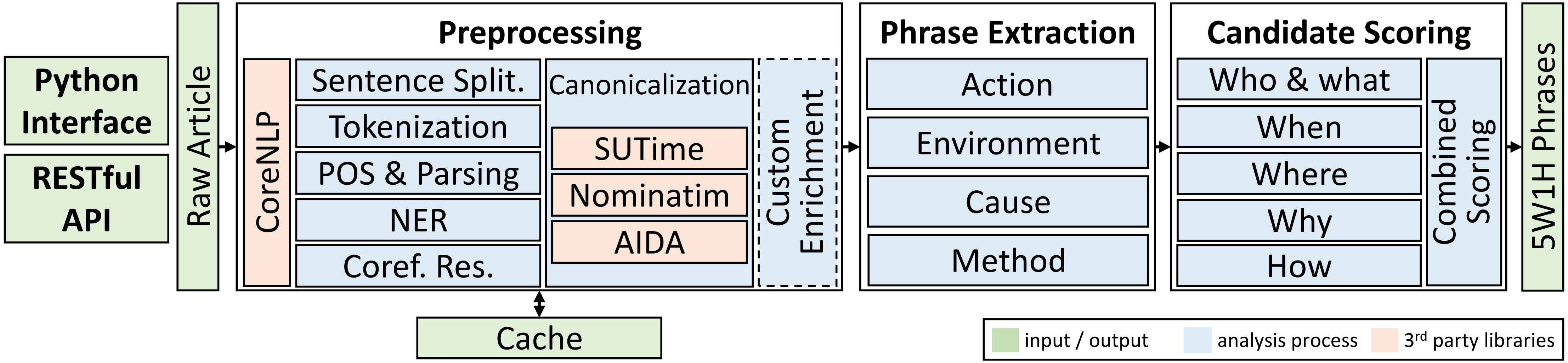
Methodology

Giveme5W1H

Three Phase Analysis Pipeline

www.isg.uni.kn

- Syntactic and domain-specific rules for extraction and scoring



Phrase Extraction

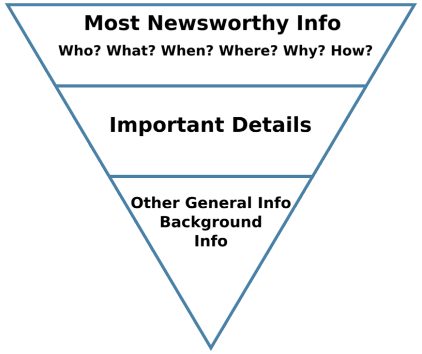
www.isg.uni.kn

- **Who**
  - **Subjects**
    (1st noun phrase (NP) in sentence)
- **What**
  - **Predicates** (verb phrase (VP) that is right to 'who' in parse tree)
- **Where**
  - Named entities (NEs) of type **location**, parsed by Nominatim
- **When**
  - NEs of type **date or time**, parsed by SUTime
- **Why**
  - Causal conjunctions (CC, "due to"), causative Vs and RBs ("implicate")
- **How**
  - Copulative CCs ("after [the train came off]"), fallback: ADJs, RBs

Candidate Scoring: Who and What

www.isg.uni.kn

- **Early** – inverted pyramid [10], but may contain hooks
- **Often**
- **Contain NE** [12]



- $s_{\text{who}}(c) = w_0\left(1 - \frac{n_{\text{pos}}(c)}{d_{\text{len}}}\right) + w_1\left(\frac{n_f(c)}{\max_{c' \in C}(n_f(c'))}\right) + w_2 \text{NE}(c)$
- $w_0 = 0.9,\ w_1 = 0.095,\ w_2 = 0.005$

- What: score jointly with respective who candidate
- Learned model parameters on 100 annotated articles

Evaluation and results

Survey Setup

www.isg.uni.kn

- Random sample of 120 articles from BBC corpus (2,225 articles) [14]
  - 24 articles for each category
    - business (Bus), entertainment (Ent), politics (Pol), sport (Spo), and tech (Tec)

- Three assessors (graduate IT students)

- 3-point Likert scale
  - Non-relevant: 0
  - Partially relevant: 0.5
  - Relevant: 1

Precision = 0.73 (on 4W: 0.82)

www.isg.uni.kn

Property	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.	.91	.72	.72	.75	.73	.71	**.73**
Avg 4W	.91	.80	.78	.86	.85	.80	**.82**

Comparison to State-of-the-Art

www.isg.uni.kn

- Only on 5W evaluated, Giveme5W1H is
  - 0.05 better than Giveme5W [17] (0.70)
  - 0.10 better than fraction of “correct” answers in [29] (0.65)
  - 0.14 worse than precision in [36] (0.89)

- **However**
  - No gold standard & use of non-disclosed datasets [29, 35, 36]
  - Input translated from other languages [29]
  - Binary relevance assessments [20,36]

Conclusion

First Open-Source 5W1H Extractor

www.isg.uni.kn

- Syntactic and domain-specific rules
- Precision = 0.73
  - Only on 4W: 0.82

- Get it at github.com/fhamborg/Giveme5W1H

open source initiative

- Future work
  - Improve “what” by scoring more independently from “who”
  - Extract implicit locations, e.g., “Apple HQ” → Cupertino

Thank You!

Web: isg.uni-konstanz.de
Mail: felix.hamborg@uni-konstanz.de