Nlp Term Project

Apoorva Kumar
Gaurav Jha
Harshad Gavali

What's Task?

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 [...].

Figure 1: News article [1] consisting of title (bold), lead paragraph (italic), and first of remaining paragraphs. Highlighted phrases represent the 5W1H event properties (who did what, when, where, why, and how).

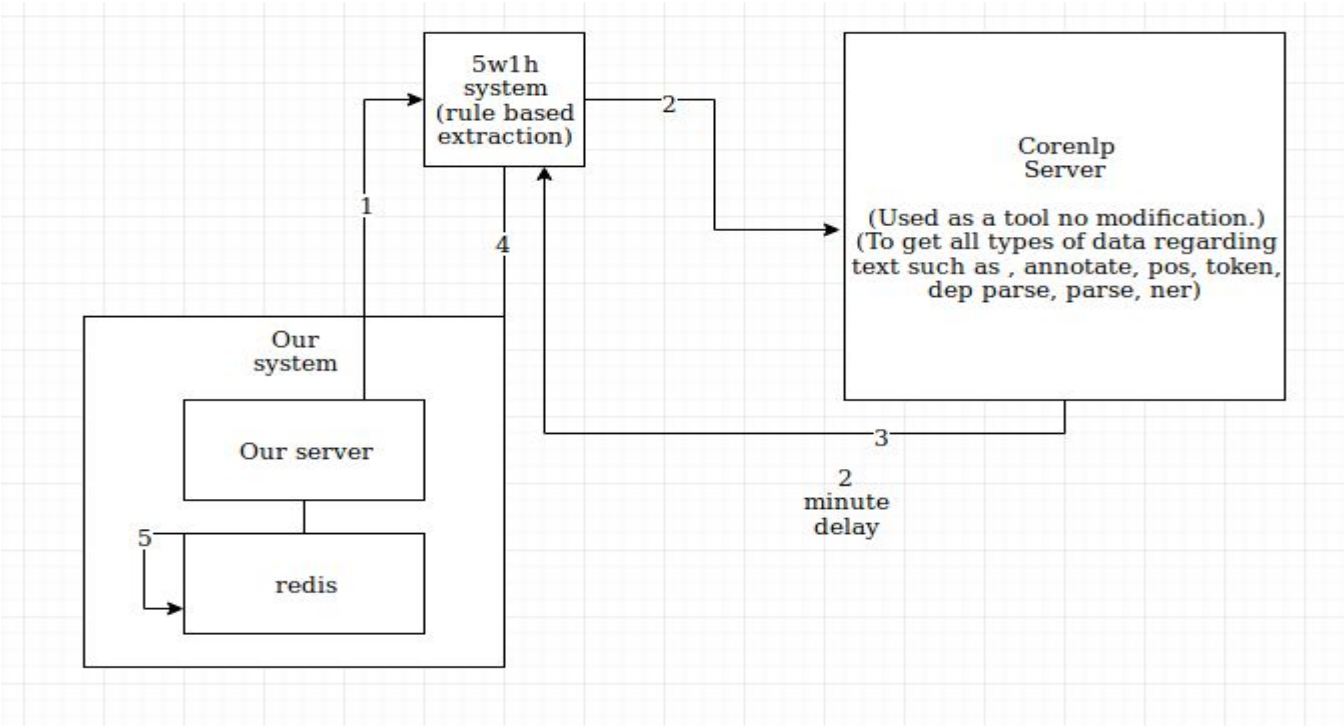
What do we know?

- The journalistic 5W1H questions are capable of describing the main event of an article
  - by answering who did what, when, where, why, and how
- Explicit event descriptors are properties that occur in a text to describe an event
  - the phrases in an article that enable a reader to understand what the article is reporting on

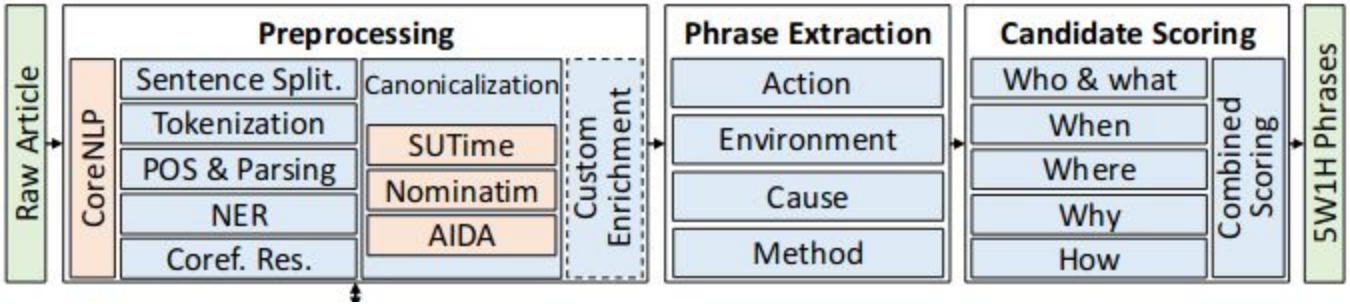
Event descriptors

- Main event descriptors is defined to be concise
  - This means they must be as short as possible
  - contain only the information describing the event
  - as long as necessary to contain all information of the even

Overview of the system



Overview of 5w1h Extraction system



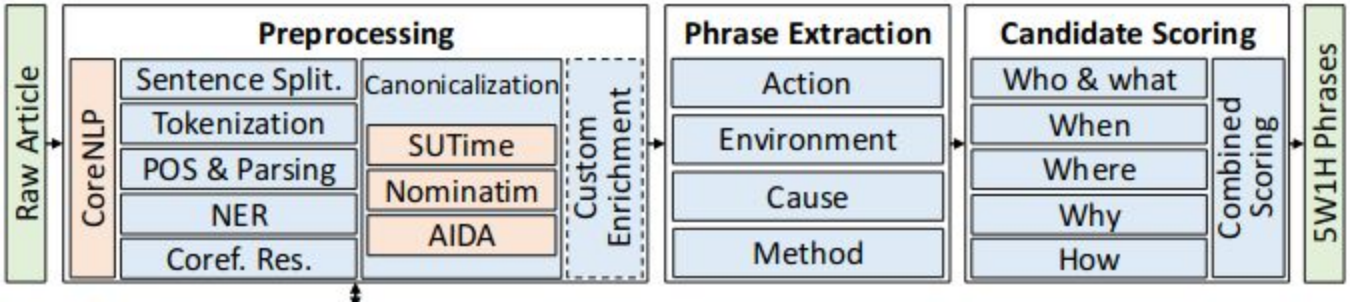
Preprocessing

- Input: text of a news article
  - including headline, lead paragraph, and body text as a one or seperate
  - Publishing date (optional)
    - which helps parse relative dates, such as “yesterday at 1 pm”.
- Stanford CoreNLP
  - sentence splitting, tokenization, lemmatization, POS-tagging, full parsing, NER and pronominal and nominal coreference resolution.

Preprocessing

- bring all NEs in the text into their canonical form
  - SUTime -> date, time -> TIMEX3 instance
    - NEs of the type date or time and merges adjacent tokens to phrases
  - Geocoding -> places and addresses to canonical geocodes,
    - coordinates referring to a point or area on earth
    - tokens classified as NEs of the type location
      - merge adjacent tokens of the same NE type within the same sentence constituent
  - Link remaining NE link to concepts in the YAGO graph

Overview of 5w1h Extraction system



Phrase extraction

- Four independent extraction chains to retrieve the article’s main event
  - the action chain extracts phrases for the ‘who’ and ‘what’ questions,
  - environment for ‘when’ and ‘where’
  - cause for ‘why’
  - method for ‘how’

Phrase extraction: Action

- Who did what?
- Who
  - subject of each sentence
- What
  - Verb Phrase that is the next sibling to each who candidate

Phrase extraction: Environment

- Retrieve phrases describing the temporal and locality context of the event.
- When
  - TIMEX3 instances from preprocessing
- Where
  - Geocodes

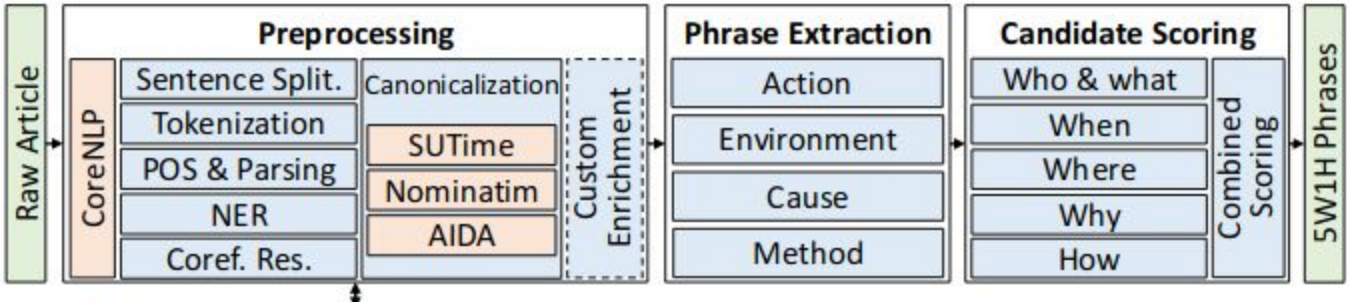
Phrase extraction; cause

- Linguistic features indicating a causal relation within a sentence constituents.
- We look for three types of cause-effect indicators
  - causal conjunctions
    - “due to”, “result of”, and “effect of”, connect two clauses, whereas the second clause yields the ‘why’ candidate
  - causative adverbs
    - “therefore”, “hence”, and “thus”, the first clause yields the ‘why’ candidate
  - causative verbs
    - “activate” and “implicate”
    - NP-VP-NP pattern, the last NP yields the ‘why’ candidate
    - Is VP causative?
      - Extract the VP’s verb, retrieve the verb’s synonyms from WordNet and compare the verb and its synonyms with the list of causative verbs ,

Phrase extraction: Method

- ‘How’ phrases
- Often, sentences with a copulative conjunction contain a method phrase in the clause that follows the copulative conjunction,
  - e.g., “after [the train came off the tracks]”.
  - Look for copulative conjunctions, If a token matches, we take the right clause as the ‘how’ candidate.
- Often, phrases that consist purely of adjectives or adverbs represent how an action was performed.
  - We use this extraction method as a fallback, copulative conjunction-based extraction too restrictive in many cases.

Overview of 5w1h Extraction system



Candidate Scoring

- Independently for each of the 5W1H questions.
- Then, adjust scores of candidates of one question dependent on properties, e.g., position, of candidates of other questions.
- For each question `q`, weighted sum of `n` scoring factors
    - score(q) = sum(w(q') * score(q'))

Candidate Scoring

- Who
  - Early : Inverse pyramid
    - News mention the most important information, i.e., the main event, early in the article
  - Often
    - Primary actor involved in the main event is likely to be mentioned frequently in the article
  - Is Named Entity
    - In news, the actors involved in events are often NEs
- What
  - Due to the strong relation between agent and action, we rank VPs according to their NPs' scores.
  - Hence, the most likely VP is the sibling in the parse tree of the most likely NP
    - score(what) = score(who)

Candidate Scoring

- When
  - Early, often,
  - Should also be close to the publishing date of the article
  - And of Relatively short duration
- Where
  - Early and often
  - It should also be often geographically contained in other location candidates
    - Geographically contained
    - Another measure of ofteness
  - And specific (Area)
    - Delhi over India

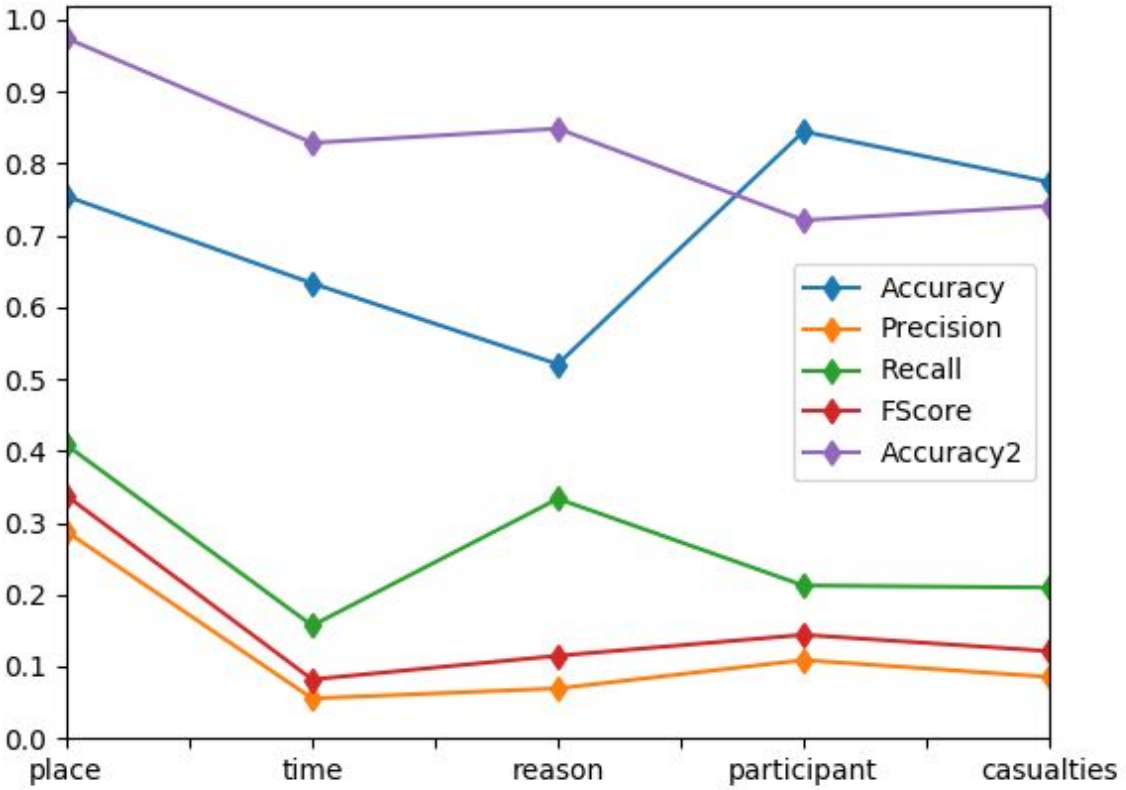
Candidate Scoring

- Why
  - Occur early in the document
  - Their causal type shall be reliable
    - Rewards causal types with low ambiguity
    - e.g., “because” has a very high likelihood that the subsequent phrase contains a cause
- Method
  - Early and often
  - Method type shall be reliable

Candidate Scoring

- The final sub-task in candidate scoring is combined scoring
  - Adjusts scores of candidates of a single 5W1H question depending on the candidates of other questions.
- A combined sentence-distance scorer
  - The assumption is that the method of performing an action should be close to the mention of the action

Result


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
place
time
reason
participant
casualties
Accuracy
Precision
Recall
FScore
Accuracy2

Result

	Accuracy%	Precision	Recall	FScore
Place	93.13	25.87	32.26	28.71
Time	97.42	28.77	40.81	33.75
Reason	82.87	5.54	15.68	8.19
Participant	84.84	6.93	33.33	11.48
Casualties	72.09	10.88	21.26	14.39
After-effects	74.08	8.53	20.98	12.13

Confusion Matrix for Place-argument

Predicted/Actual	1	0
1	2435	6977
0	5112	161686

Confusion Matrix for Time Argument

Predicted/Actual	1	0
1	1088	2693
0	1578	160504

Confusion Matrix for Reason argument

Predicted/Actual	1	0
1	620	10563
0	3334	66625

Confusion Matrix for Participant Argument

Predicted/Actual	1	0
1	282	3783
0	564	24064

Confusion Matrix for Casualties argument

Predicted/Actual	1	0
1	3963	32450
0	14676	117813

Confusion Matrix for After-Effect argument

Predicted/Actual	1	0
1	830	8895
0	3126	33527

Merits and Demerits

- Merits
  - It provides decent accuracy for all the expected arguments.
  - Easy to setup and run.
- Demerits
  - ‘Where’ only matches Named Entity of type Location
    - “The fire caught in the building” it won’t give “building” as where argument.
  - False positive is high, this is due to the fact that data is not properly annotated and we are mapping one or two arguments from the same arguments provided by our system.

Future Work

- Joint extraction of optimal ‘who’ candidates with non-optimal ‘what’ candidates and cut-off what candidates
  - the headline contained a concise ‘who’ phrase but the ‘what’ phrase did not contain all information
    - e.g., because it only aimed to catch the reader’s interest, a journalistic hook
  - Devise separate extraction methods for both questions.
    - Thereby, we need to ensure that the top candidates of both questions fit to each other
- The date of an event may be implicitly defined by the reported event
  - e.g., “in the final of the Canberra Classic”.
- The location may be implicitly defined by the main actor,
  - e.g., “Apple Postpones Release of [...]”, which likely happened at the Apple headquarters in Cupertino.
  - Similarly, the proper noun “Stanford University” also defines a location

Thank You.