Time Expression Analysis and Recognition Using Syntactic Token Types and General Heuristic Rules

Xiaoshi Zhong and Erik Cambria
Computer Science and Engineering
Nanyang Technological University
{xszhong, cambria}@ntu.edu.sg

Outline

- Time expression analysis
    - Datasets: TimeBank, Gigaword, WikiWars, Tweets
    - Findings: short expression, occurrence, small vocabulary, similar syntactic behavior
- Time expression recognition
    - SynTime: syntactic token types and general heuristic rules
    - Baselines: HeidelTime, SUTime, UWTime

Time Expression Analysis

- Datasets
  - TimeBank
  - Gigaword
  - WikiWars
  - Tweets
- Findings
  - Short expression
  - Occurrence
  - Small vocabulary
  - Similar syntactic behaviour

Example time expressions:

now
today
Friday
February
the last week
13 January 1951
June 30, 1990
8 to 20 days
the third quarter of 1984
…

Time Expression Analysis - Datasets

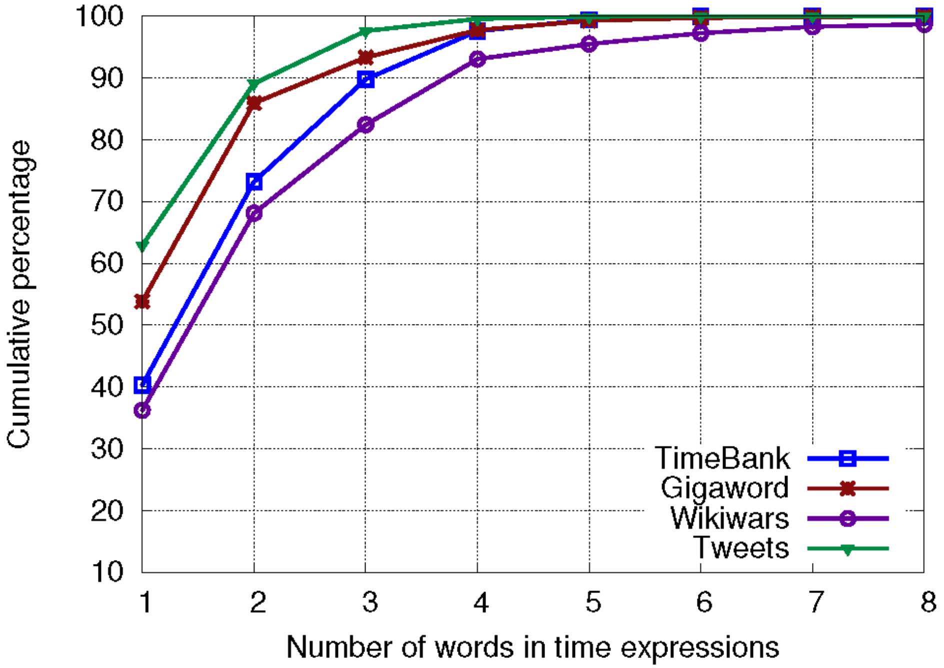
- Datasets
  - TimeBank: a benchmark dataset used in TempEval series
  - Gigaword: a large dataset with generated labels and used in TempEval-3
  - WikiWars: a specific domain dataset collected from Wikipedia about war
  - Tweets: a manually labeled dataset with informal text collected from Twitter
- Statistics of the datasets

Dataset	**#Docs**	**#Words**	**#TIMEX**
TimeBank	183	61,418	1,243
Gigaword	2,452	666,309	12,739
WikiWars	22	119,468	2,671
Tweets	942	18,199	1,127

The four datasets vary in source, size, domain, and text type, but we will see that their time expressions demonstrate similar characteristics.

Time Expression Analysis – Finding 1

- **Short expression**: time expressions are very short.



Time expressions follow a similar length distribution

80% of time expressions contain ⩽3 words
90% of time expressions contain ⩽4 words

Average length of time expressions

Dataset	**Average length**
TimeBank	2.00
Gigaword	1.70
WikiWars	2.38
Tweets	1.51

Average length: about 2 words

Time Expression Analysis – Finding 2

- **Occurrence**: most of time expressions contain time token(s).

Percentage of time expressions that contain time token(s)

Dataset	Percentage
TimeBank	94.61
Gigaword	96.44
WikiWars	91.81
Tweets	96.01

Example time tokens (red):

now
today
Friday
February
the last week
13 January 1951
June 30, 1990
8 to 20 days
the third quarter of 1984
…

Time Expression Analysis – Finding 3

- **Small vocabulary**: only a small group of time words are used to express time information.

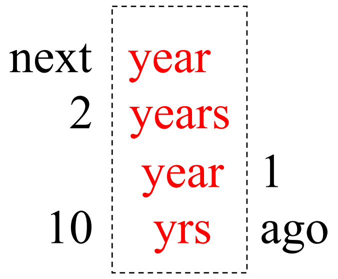
Number of **distinct** words and time tokens in time expressions

Dataset	#Words	#Time tokens
TimeBank	130	64
Gigaword	214	80
WikiWars	224	74
Tweets	107	64

Number of distinct words and time tokens **across** four datasets

	#Words	#Time tokens
	350	123

45 distinct time tokens appear in all the four datasets.
That means, time expressions highly overlap at their time tokens.


next year
2 years
year 1
10 yrs ago

Overlap at year

Time Expression Analysis – Finding 4

- **Similar syntactic behavior**: (1) POS information cannot distinguish time expressions from common text, but (2) within time expressions, POS tags can help distinguish their constituents.
  - (1) For the top 40 POS tags (10 × 4 datasets), 37 have percentage lower than 20%, other 3 are CD.
  - (2) Time tokens mainly have NN* and RB, modifiers have JJ and RB, and numerals have CD.

Time Expression Analysis – Eureka!

- **Similar syntactic behavior**: (1) POS information cannot distinguish time expressions from common text, but (2) within time expressions, POS tags can help distinguish their constituents.
    - (1) For the top 40 POS tags (10 × 4 datasets), 37 have percentage lower than 20%, other 3 are CD.
    - **(2) Time tokens mainly have NN* and RB, modifiers have JJ and RB, and numerals have CD.**

When seeing (2), we realize that this is exactly how linguists define part-of-speech for language; similar words have similar syntactic behaviour. The definition of part-of-speech for language inspires us to define a type system for the time expression, part of language.

Our Eureka! moment

Time Expression Analysis - Summary

- Summary
  - On average, a time expression contains two tokens; one is time token and the other is modifier/numeral. And the time tokens are in small size.
- Idea for recognition
  - To recognize a time expression, we first recognize the time token, then recognize the modifier/numeral.

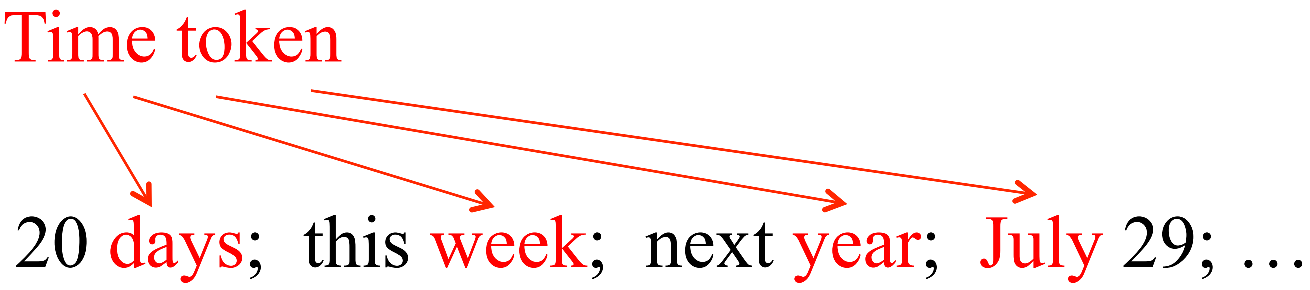
Time Expression Analysis - Idea

- Summary
  - On average, a time expression contains two tokens; one is time token and the other is modifier/numeral. And the time tokens are in small size.
- Idea for recognition
  - To recognize a time expression, we first recognize the time token, then recognize the modifier/numeral.

20 days; this week; next year; July 29; …

Time Expression Analysis - Idea

- Summary
  - On average, a time expression contains two tokens; one is time token and the other is modifier/numeral. And the time tokens are in small size.
- Idea for recognition
  - To recognize a time expression, we first recognize the time token, then recognize the modifier/numeral.

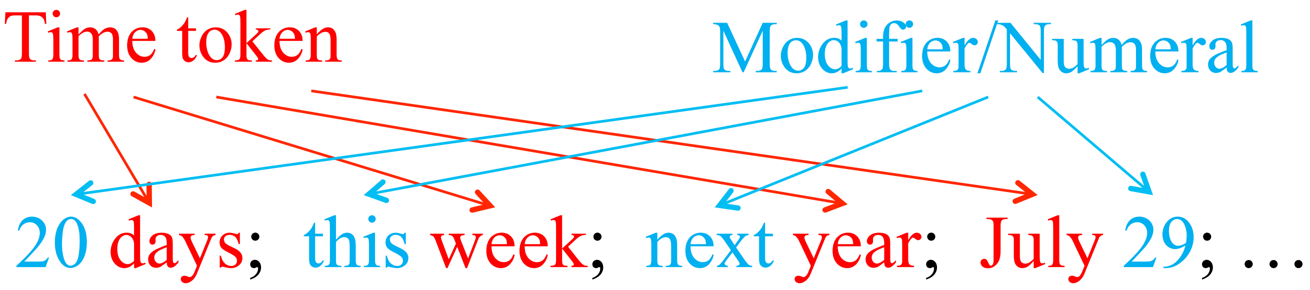


Time token

20 days; this week; next year; July 29; …

Time Expression Analysis - Idea

- Summary
  - On average, a time expression contains two tokens; one is time token and the other is modifier/numeral. And the time tokens are in small size.
- Idea for recognition
  - To recognize a time expression, we first recognize the time token, then recognize the modifier/numeral.



Time token

Modifier/Numeral

20 days; this week; next year; July 29; …

Time Expression Recognition

- SynTime
  - Syntactic token types
  - General heuristic rules
- Baseline methods
  - HeidelTime
  - SUTime
  - UWTime
- Experiment datasets
  - TimeBank
  - WikiWars
  - Tweets

Time Expression Recognition - SynTime

- Syntactic token types
- General heuristic rules

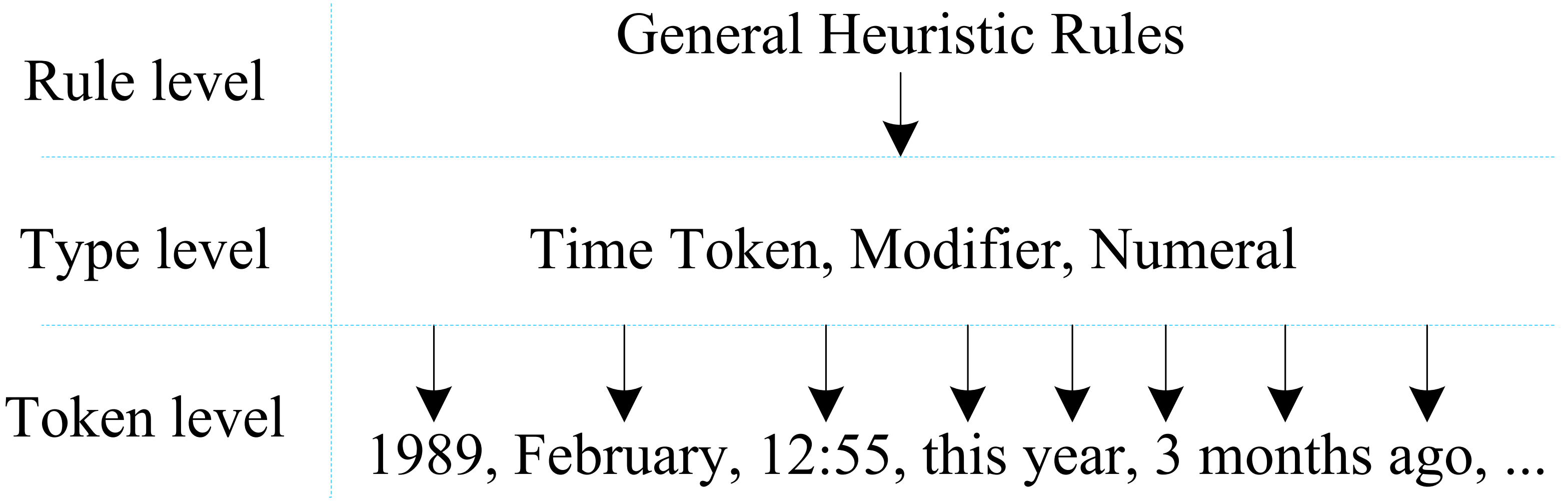
Time Expression Recognition - SynTime

- Syntactic token types – A type system
  - Time token: explicitly express time information, e.g., “year”
    - 15 token types: DECADE, YEAR, SEASON, MONTH, WEEK, DATE, TIME, DAY_TIME, TIMELINE, HOLIDAY, PERIOD, DURATION, TIME_UNIT, TIME_ZONE, ERA
  - Modifier: modify time tokens, e.g., “next” modifies “year” in “next year”
    - 5 token types: PREFIX, SUFFIX, LINKAGE, COMMA, IN_ARTICLE
  - Numeral: ordinals and numbers, e.g., “10” in “next 10 years”
    - 1 token type: NUMERAL
  - **Token types to tokens is like POS tags to words**
    - POS tags: next/JJ 10/CD years/NNS
    - Token types: next/PREFIX 10/NUMERAL years/TIME_UNIT

Time Expression Recognition - SynTime

- General heuristic rules
  - Only relevant to token types
  - Independent of specific tokens

SynTime – Layout



Rule level

General Heuristic Rules

Type level

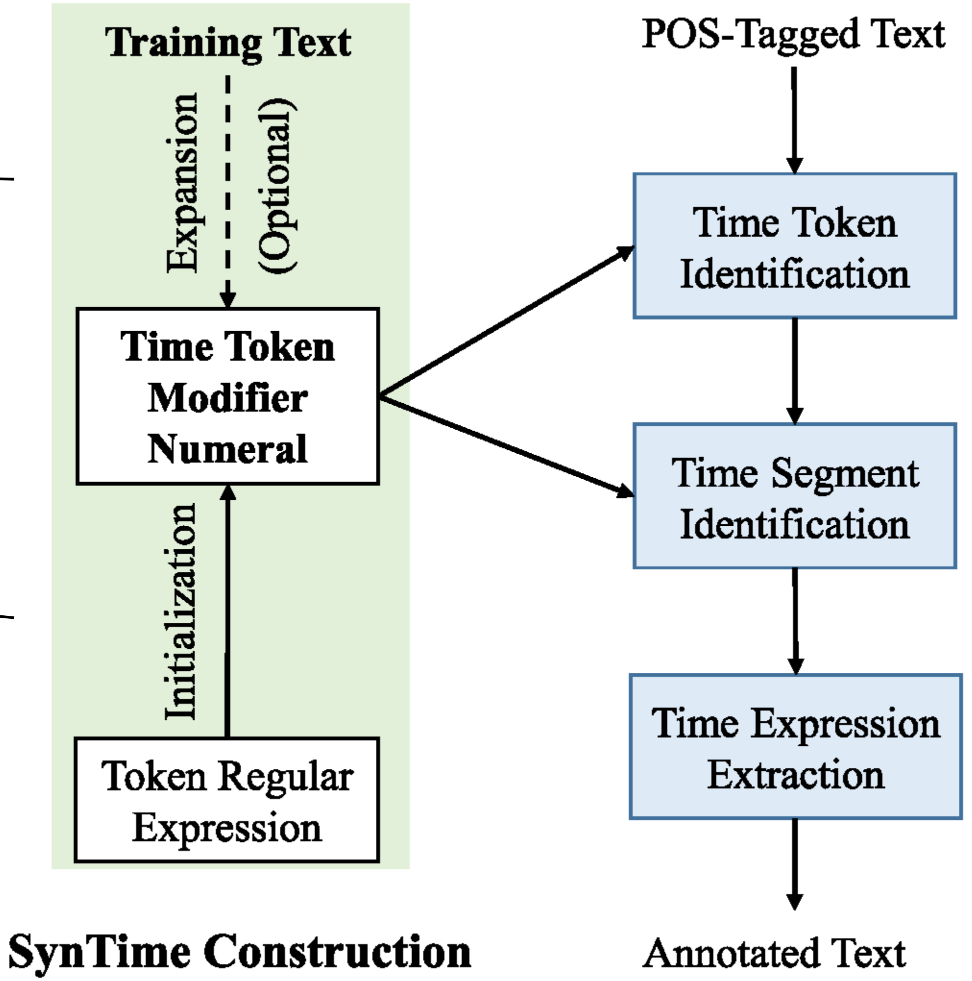
Time Token, Modifier, Numeral

Token level

1989, February, 12:55, this year, 3 months ago, ...

Token level: time-related tokens and token regular expressions
Type level: token types group the tokens and token regular expressions
Rule level: heuristic rules work on token types and are independent of specific tokens

SynTime – Overview in practice



Training Text

Expansion (Optional)

Time Token
Modifier
Numeral

Initialization

Token Regular Expression

SynTime Construction

POS-Tagged Text

Time Token Identification

Time Segment Identification

Time Expression Extraction

Annotated Text

Add keywords under defined token types and do not change any rules

Import token regex to time token, modifier, numeral

Identify time tokens

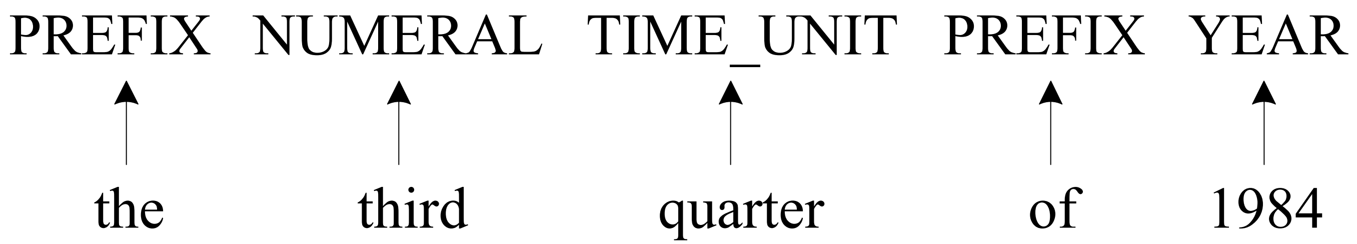
Identify modifiers and numerals by expanding the time tokens’ boundaries

Extract time expressions

An example: the third quarter of 1984

A sequence of tokens: the third quarter of 1984

An example: the third quarter of 1984



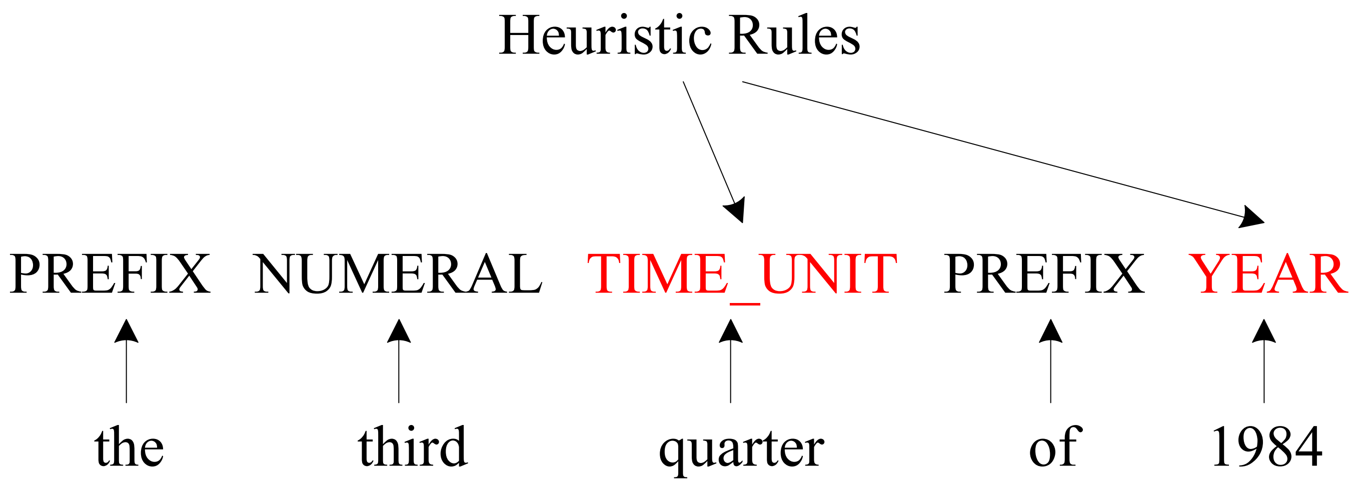
Assign tokens with token types

PREFIX NUMERAL TIME_UNIT PREFIX YEAR

A sequence of tokens:

the third quarter of 1984

An example: the third quarter of 1984



Heuristic Rules

Identify time tokens

Assign tokens with token types

PREFIX NUMERAL TIME_UNIT PREFIX YEAR

A sequence of tokens:

the third quarter of 1984

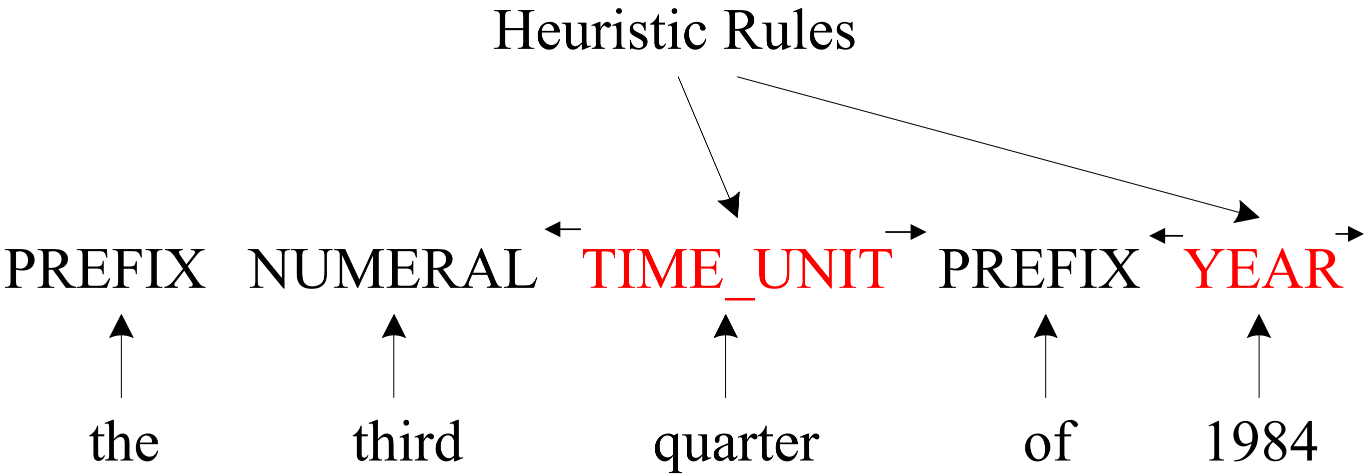
An example: the third quarter of 1984

Identify modifiers and numerals by searching time tokens' surroundings

Identify time tokens

Assign tokens with token types

A sequence of tokens:



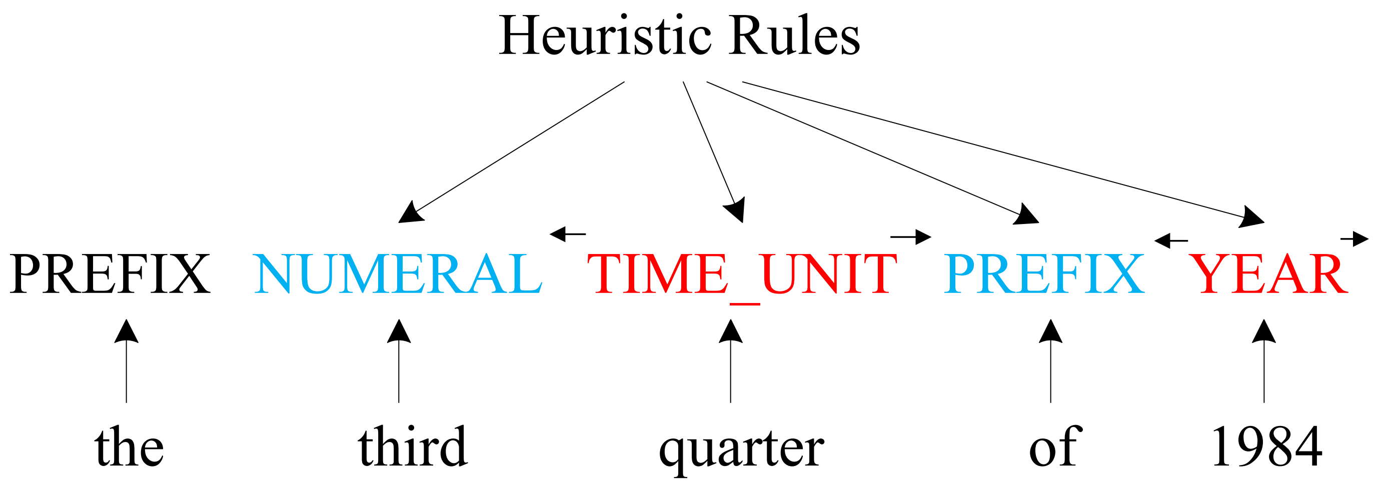
An example: the third quarter of 1984

Identify modifiers and numerals by searching time tokens' surroundings

Identify time tokens

Assign tokens with token types

A sequence of tokens:



Heuristic Rules

PREFIX NUMERAL TIME_UNIT PREFIX YEAR

the third quarter of 1984

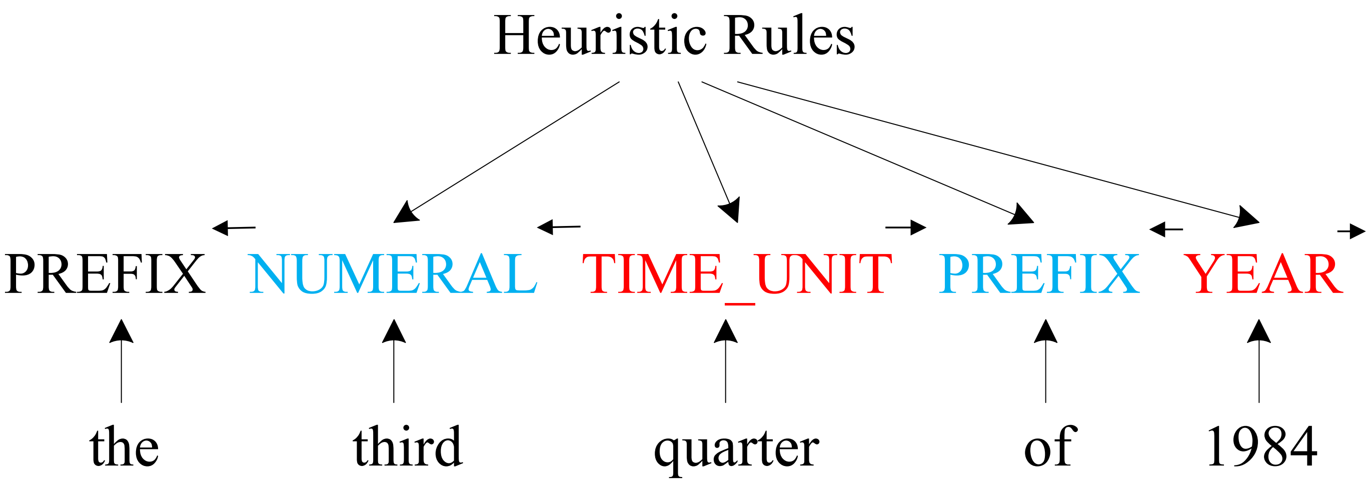
An example: the third quarter of 1984

Identify modifiers and numerals by searching time tokens’ surroundings

Identify time tokens

Assign tokens with token types

A sequence of tokens:



Heuristic Rules

PREFIX NUMERAL TIME_UNIT PREFIX YEAR

the third quarter of 1984

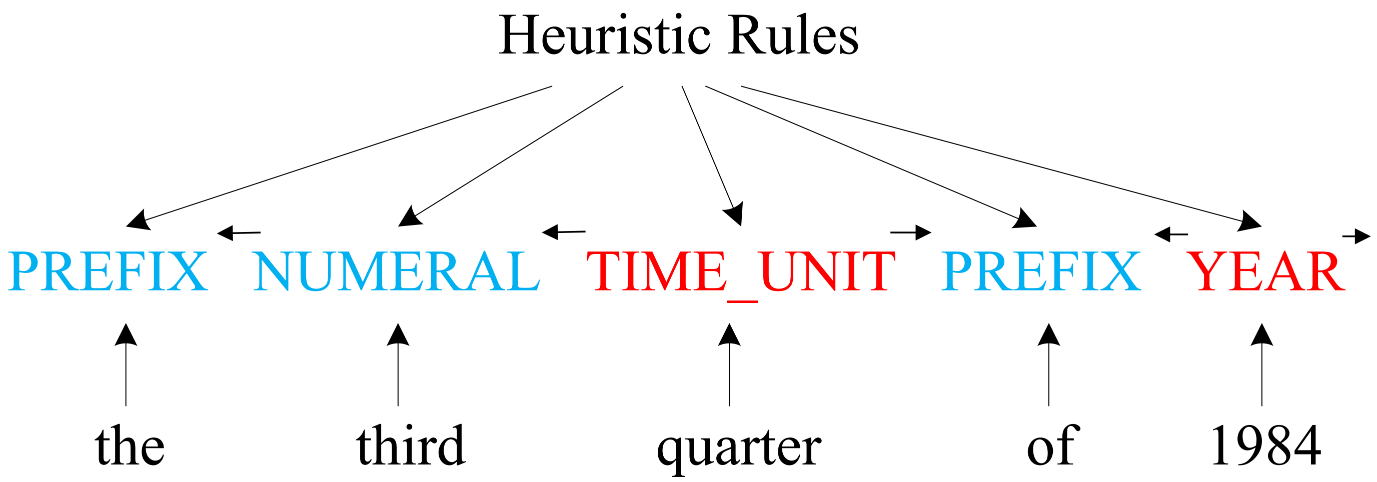
An example: the third quarter of 1984

Identify modifiers and numerals by searching time tokens' surroundings

Identify time tokens

Assign tokens with token types

A sequence of tokens:



Heuristic Rules

PREFIX NUMERAL TIME_UNIT PREFIX YEAR

the third quarter of 1984

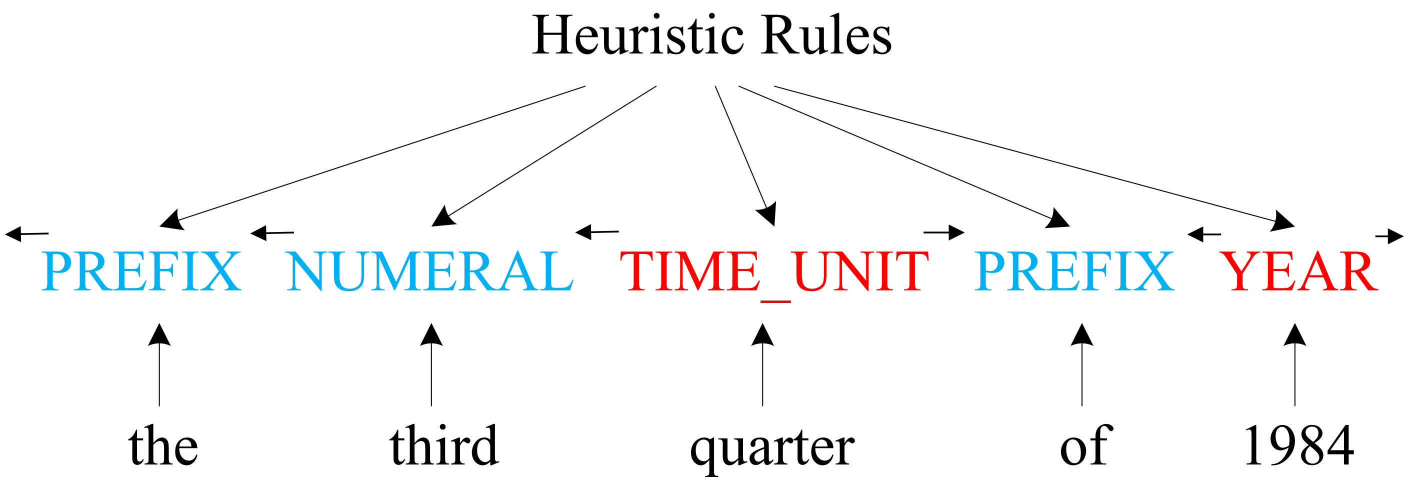
An example: the third quarter of 1984

Identify modifiers and numerals by searching time tokens' surroundings

Identify time tokens

Assign tokens with token types

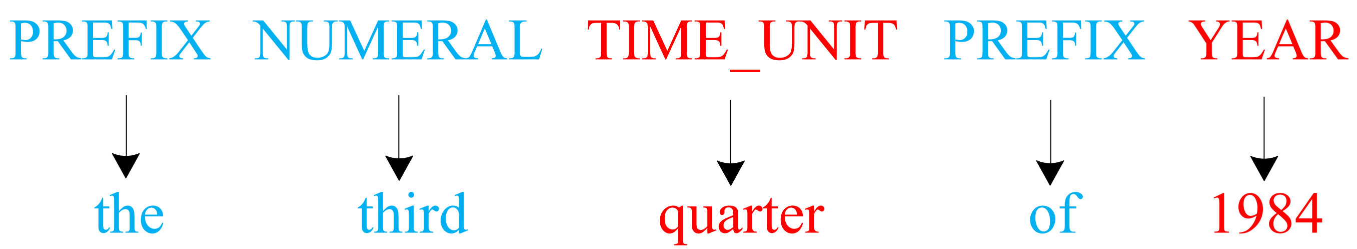
A sequence of tokens:



An example: the third quarter of 1984

A sequence of token types PREFIX NUMERAL TIME_UNIT PREFIX YEAR

An example: the third quarter of 1984



A sequence of token types

PREFIX NUMERAL TIME_UNIT PREFIX YEAR

Export a sequence of tokens
as time expression

the third quarter of 1984

An example: the third quarter of 1984

Time expression: the third quarter of 1984

Time Expression Recognition - Experiments

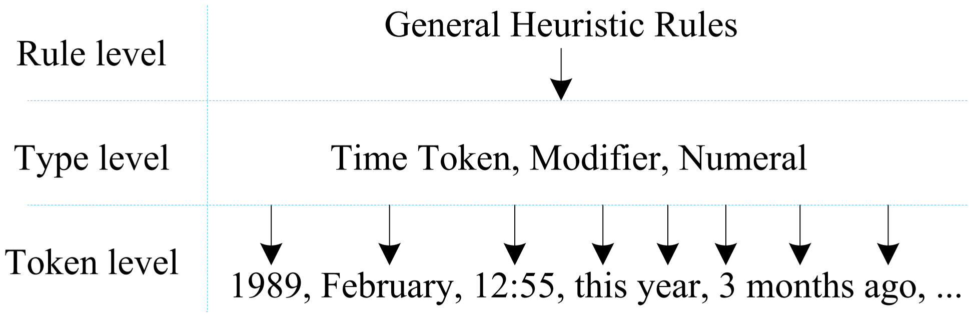
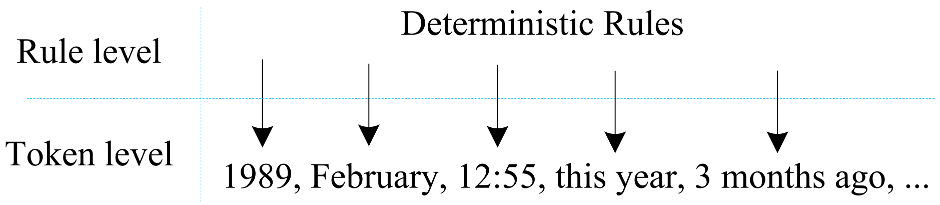
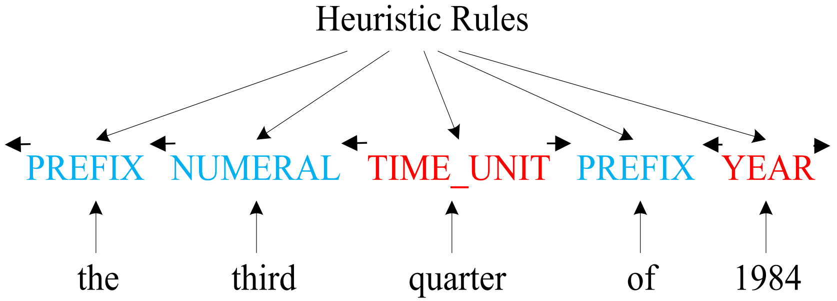
- SynTime
  - SynTime-I: **I**nitial version
  - SynTime-E: **E**xpanded version, adding keywords to SynTime-I

  (Add keywords under the defined token types and do not change any rules.)
- Baseline methods
  - HeidelTime: rule-based method
  - SUTime: rule-based method
  - UWTime: learning-based method
- Experiment datasets
  - TimeBank: comprehensive data in formal text
  - WikiWars: specific domain data in formal text
  - Tweets: comprehensive data in informal text

Overall performance. The **best results** are in boldface and the second best are underlined. Some results are borrowed from their original papers and the papers indicated by the references.

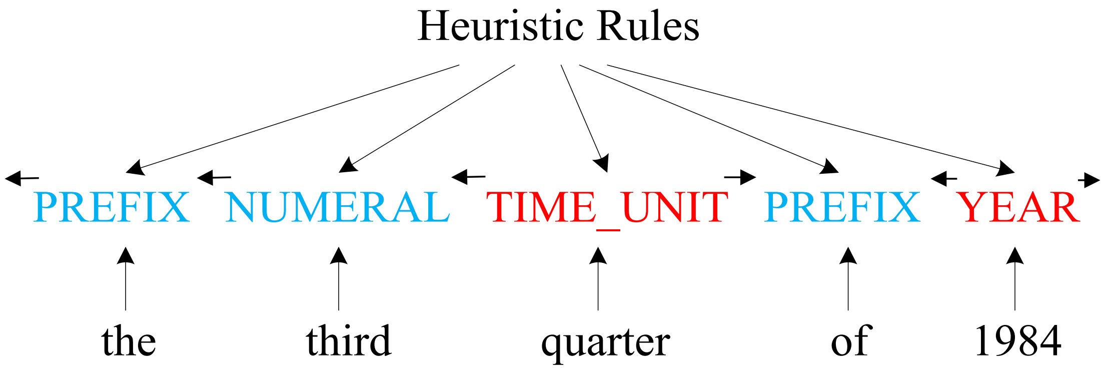
Dataset	Methods	Strict Match			Relexed Match		
		Pr.	*Re.*	*F1*	*Pr.*	*Re.*	*F1*
TimeBank	HeidelTime(Strotgen et al., 2013)	83.85	78.99	81.34	93.08	87.68	90.30
	SUTime(Chang and Manning, 2013)	78.72	80.43	79.57	89.36	91.30	90.32
	UWTime(Lee et al., 2014)	86.10	80.40	83.10	**94.60**	88.40	91.40
	SynTime-I	91.43	92.75	92.09	94.29	**95.65**	**94.96**
	SynTime-E	**91.49**	**93.48**	**92.47**	93.62	**95.65**	94.62
WikiWars	HeidelTime(Lee et al., 2014)	85.20	79.30	82.10	92.60	86.20	89.30
	SUTime	78.61	76.69	76.64	95.74	89.57	92.55
	UWTime(Lee et al., 2014)	**87.70**	78.80	**83.00**	**97.60**	87.60	92.30
	SynTime-I	80.00	80.22	80.11	92.16	92.41	92.29
	SynTime-E	79.18	**83.47**	81.27	90.49	**95.39**	**92.88**
Tweets	HeidelTime	**89.58**	72.88	80.37	95.83	77.97	85.98
	SUTime	76.03	77.97	76.99	88.43	90.68	89.54
	UWTime	88.54	72.03	79.44	**96.88**	78.81	86.92
	SynTime-I	89.52	94.07	91.74	93.55	98.31	95.87
	SynTime-E	89.20	**94.49**	**91.77**	93.20	**98.78**	**95.88**

Difference from other Rule-based Methods

Method	SynTime	Other rule-based methods
Layout	 Rule level General Heuristic Rules Type level Time Token, Modifier, Numeral Token level 1989, February, 12:55, this year, 3 months ago, ...	 Rule level Deterministic Rules Token level 1989, February, 12:55, this year, 3 months ago, ...
Property	Heuristic rules work on token types and are independent of specific tokens, thus they are independent of specific domains and specific text types and specific languages.	Deterministic rules directly work on tokens and phrases in a fixed manner, thus the taggers lack flexibility
Example	 Heuristic Rules PREFIX NUMERAL TIME_UNIT PREFIX YEAR the third quarter of 1984	/the/? [{tag:JJ}]? ($NUM_ORD) /-/? [{tag:JJ}]? /quarter/

A simple idea

Rules can be designed with generality and heuristics


Heuristic Rules
PREFIX
NUMERAL
TIME_UNIT
PREFIX
YEAR
the
third
quarter
of
1984