



A Probabilistic Model with Commonsense Constraints for Pattern-based Temporal Fact Extraction

Yang Zhou, Tong Zhao, Meng Jiang

Task: Temporal Fact Extraction from unstructured text data

w.r.t finding Country's President Person at Time Task

- 1) "... The former French president Jacques Chirac, a self-styled affable rogue who was head of state from 1995 to 2007 ..." (posted on Sept. 26, 2019 [text gen. time]) "
- 2) "... Emmanuel Macron, now President of France, graduated from ENA in 2004..." (posted on Sept. 19, 2019) "



Extracted by

--From news data

Textual Pattern

- Pattern 1: former Country president Person
- Pattern 2: Person, now president of Country

Time Signal:

- temporal tag
- text generate time

Temporal Fact Extraction from unstructured text data

1) “... The former **French** [Country: France] president **Jacques Chirac** [Person], a self-styled affable rogue who was head of state from **1995** [temporal tag] to 2007 ...” (posted on Sept. 26, **2019** [text gen. time]) “

2) “... **Emmanuel Macron** [Person], now President of **France** [Country], graduated from ENA in **2004** [temporal tag] ...” (posted on Sept. 19, **2019** [text gen. time]) “

--From news data

Textual Pattern

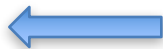
Pattern 1: former Country president Person

Pattern 2: Person, now president of Country

Time Signal:

- temporal tag
- text generate time

Temporal Fact Extraction



- ✓ (France, Jacques Chirac, 1995): **P1** and **temporal tag**;
- ✗ (France, Jacques Chirac, 2019): **P1** and **text gen.time**;
- ✗ (France, Emmanuel Macron, 2004): **P2** and **temporal tag**;
- ✓ (France, Emmanuel Macron, 2019): **P2** and **text gen.time**.

Here, we have some observations about Temporal Fact Extraction:

O1: Not every pattern is reliable, Patterns have reliability .

pattern such as “*Person visited Country*” is very likely to be unreliable; and pattern such as “*current Country’s president Person*” is very likely to be reliable.

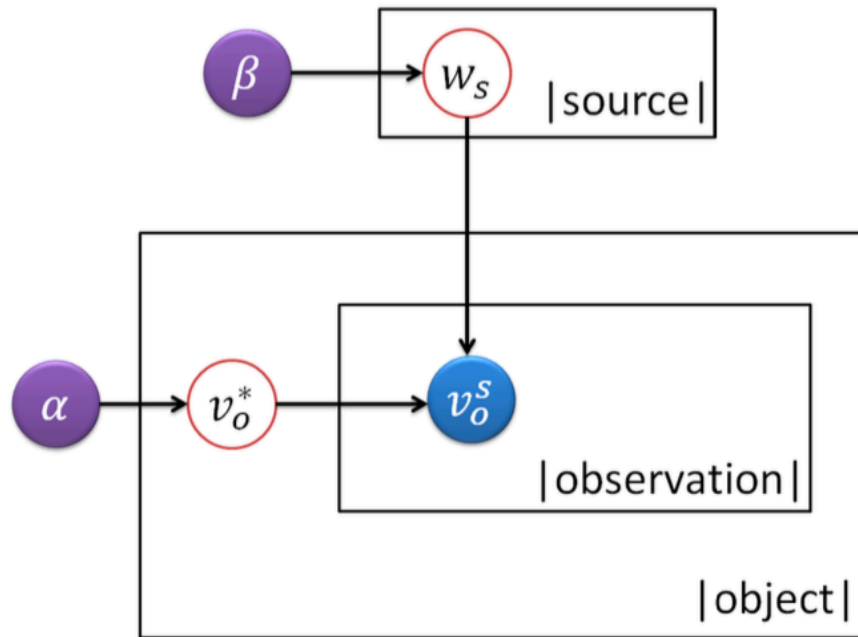
O2: There is a dependency between pattern and type of time signal

For temporal fact extraction, different types of time signals might be either reliable or unreliable depending on the pattern.

Truth Discovery via PGM

Truth discovery approaches follow two fundamental principles:

- (1) If a **source** provides much trustworthy **information**, its reliability is high
- (2) If an **Information** is supported by many reliable **source**, this **information** is more likely to be true.



How to design a PGM for temporal truth discovery? ?

Temporal Truth: Commonsense constraint

Here, we have some **commonsense constraint** about Temporal Truth:

For country's president:

- one president serves only one country;
- one country has only one president at a time;
- one country can have multiple presidents in the history (e.g., United States, France).

For sports team's player:

- one player serves only one club at a time;
- one club has multiple players and one player can serve multiple clubs in his/her career.



generalize

- C1: one value matches with only one entity;
- C2: one entity matches with only one value;
- C3: one value matches with only one entity at a time;
- C4: one entity matches with only one value at a time.

Temporal Truth: Commonsense constraint

Commonsense Constraint Rules

- C1: one value matches with only one entity;
- C2: one entity matches with only one value;
- C3: one value matches with only one entity at a time;
- C4: one entity matches with only one value at a time.

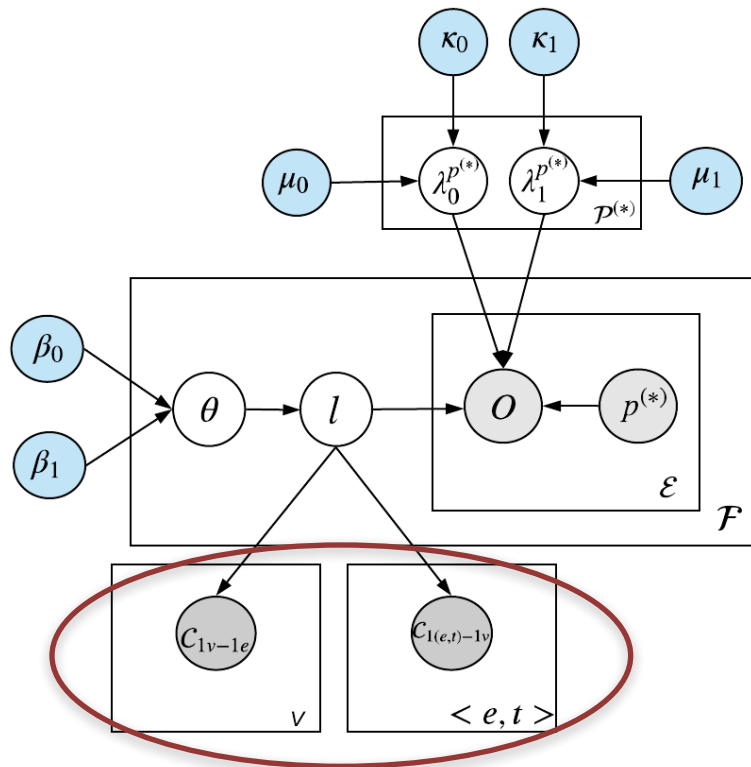
However, In probabilistic graphic model, all the nodes are variable

How to add Commonsense Constraint Rules to PGMs ?



PGMCC:

Probabilistic Graphical Model with Commonsense Constraint



Constraint variable

Symbol	Description
θ_f	$[0, 1]$, trustworthiness of temporal fact tuple f
l_f	Boolean: label of temporal fact f
o_e	Integer: the observed frequency of fact f_e extracted by pattern $p_e^{(*)}$
$\lambda_0^{p^{(*)}}, \lambda_1^{p^{(*)}}$	Real numbers: reliability of pattern $p^{(*)}$ on giving false/true fact tuples
C_{1v-1e}	Real number: the number of entities given one value v
$C_{1(e,t)-1v}$	Real number: the sum of values given one entity e and one time t
Hyper-Parameter	
μ_0, μ_1	Integers: prior counts of false/true tuples extracted by a textual pattern
κ_0, κ_1	Integers: prior sums of false/true tuples extracted by a textual pattern
β_0, β_1	Integers: prior counts of false/true tuples

Table 2: Symbols and their descriptions used in the model.

Take a **MCMC** method to inference it.

Dataset:

- 9,876,086 news articles (4 billion words) published from 1994– 2010.
- focus on attribute country's president.
- 57,472 textual patterns, 116,631 temporal fact tuples, and 1,326,164 extractions.

Experiment Result:

- Compare with Truth discovery model(without constraint) **LTM**, PGMCC improve the AUC and F1 by **40%+**.
- Compare with Truth finding method **TFWIN** (a bootstrap method not PGMs), PGMCC improve the AUC and F1 by **7%+**.

PGMCC: case study

Method	Entity e	Value v	Year t
PGMCC $C_{1(e,t)-1v}$	France	j.r_chirac	1995
	France	j.r_chirac	1996
	France	j.r_chirac	1997
	France	j.r_chirac	1998
	France	j.r_chirac (n.s_sarkozy)	1993
	Spain (France)	j.r_chirac	1996
	Greece (France)	j.r_chirac	2003
	Tunisia (France)	j.r_chirac	2003
PGMCC $C_{1(e,t)-1v}$, C_{1v-1e}	France	j.r_chirac	1995
	France	j.r_chirac	1996
	France	j.r_chirac	1999
	France	j.r_chirac	1997
	France	j.r_chirac	1998
	Spain	l_enrique	1996
	Greece	c._photopoulos (k_stephanopoulos)	2003
	Tunisia	a_ben_ali	2003

Table 4: False case analysis for comparing PGMCC of partial and complete commonsense constraints.

$C_{1(e,t)-1v} \rightarrow$ one country one year has only one president

$C_{1v-1e} \rightarrow$ one President only serve one Country

Red means false, **Green** means right answer