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

Anonymous Author(s)

ABSTRACT

Textual patterns (e.g., `Country`'s president `Person`) are specified and/or generated for extracting factual information from unstructured data. Pattern-based information extraction methods have been recognized for their efficiency and transferability. However, not every pattern is reliable; a major challenge is to derive the most complete and accurate facts from diverse and sometimes conflicting extractions. In this work, we propose a probabilistic graphical model, which makes the first attempt to formulate fact extraction in a generative process, that can automatically infer true facts and pattern reliability without any supervision. It has two novel designs specially for temporal facts: (1) it models pattern reliability on two types of time signals, including temporal tag in text and text generation time; (2) it models commonsense constraints such as *"one president serves only one country"* and *"one country only have one president at a time"* as observable variables. Experimental results demonstrate that our model significantly outperforms existing methods on extracting true temporal facts from news data.

ACM Reference Format:
Anonymous Author(s). 2020. A Probabilistic Model with Commonsense Constraints for Pattern-based Temporal Fact Extraction. In *Proceedings of the Web Conference (TheWebConf'20), April 20–24, 2020, Taipei.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313435

1 INTRODUCTION

Temporal fact extraction, which is to extract (entity, value, time)-factual tuples from text data (e.g., news, tweets) for specific attributes, acts as one of the fundamental tasks in knowledge base construction, knowledge graph population, and question answering. For example, if we were interested in *country's president*, the entity would be of type `Location.Country`, the value would be of type `Person`, and the time would be a valid year in the person's presidential term. Thanks to name entity recognition (NER) and typing systems [3, 19], pattern-based information extraction methods, specifically, the *typed* pattern methods which generate patterns consisted of entity types and words [9, 10, 18], have been recognized as one of the most effective methodologies for their advantages of (1) good transferability (i.e., they are plug-and-use methods and can be used across domains and across datasets), (2) unsupervised manner (i.e., they require no or very few annotations), and (3) high efficiency (i.e., pattern generation is much faster than training neural architectures). So, given a huge set of text data (e.g., 9.9 million news articles we have), applying typed patterns to the task of temporal fact extraction is a new and promising idea. Note that the typed patterns only tell the association between entity (e.g., country) and value (e.g., person). Two types of time signals can be attached to the pairs, forming temporal triples: One is temporal tag in text, e.g., the year tag next to the entity/value mentions in the sentence; the other is text generation time, i.e., the year this text document (e.g., news, tweet) was posted. Let us take a look at an example as follows. Given two sentences:

> 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])

pattern-based methods can discover two typed patterns:

- **P1:** former `Country` president `Person`
- **P2:** `Person`, now president of `Country`,

Then the methods can extract the following tuples. We label ✔ and ✗ for correct tuples and incorrect ones, respectively:

- ✔ (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**.

So, we have the following observations:

- **O1:** Not every pattern is reliable: the pattern "`Person` visited `Country`" is very likely to be unreliable. Not every pattern is unreliable: the pattern "current `Country`'s president `Person`" is very likely to be reliable. The above two pattern examples are somehow half and half. So, patterns have reliability.
- **O2:** For temporal fact extraction, different types of time signals might be either reliable or unreliable depending on the pattern. So, there is a dependency between pattern and type of time signal, in terms of reliability.

Existing truth finding approaches assumed that a structured "source-object-claim" database was given and then estimated the reliability of source for inferring whether the claim was true or false [27, 31, 32]. For example, a source could be a book seller, an object could be a book's author list, and a claim could be an author list that a seller gave for a book. One conclusion was that *probabilistic graphical models* (PGM) [31, 32] have advantages of estimating source reliability over the general data distributions, compared with bootstrapping algorithms [10, 24, 27]. However, PGM-based truth finding models have never been developed for the task of information extraction. Estimating the reliability of textual patterns is new (O1). Moreover, when we focus on temporal fact extraction, modeling the dependency between pattern and type of time signals brings new challenges (O2).

TheWebConf'20, April 20–24, 2020, Taipei
© 2020 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313435

Source $p^{(*)}$: (pattern, type of time signal)		Temporal fact tuple f			Observed frequency o
Country president Person	temporal tag	united_states	geoerge_w._bush	2003	422
Country president Person	temporal tag	united_states	barrack_obama	2009	339
Country president Person	temporal tag	united_states	geoerge_w._bush	1969	23
former Country president Person	text gen. time	united_states	geoerge_w._bush	2009	268
former Country president Person	text gen. time	united_states	jimmy_carter	2008	179
former Country president Person	text gen. time	united_states	william_jefferson_clinton	2009	173
...	...	...	...	...	...

Table 1: Data input for the task of extracting facts of country's president. A source has a pattern and a type of time signal. A tuple has an entity (country), a value (person), and a time (year). The observed frequency is the number of such a tuple being matched by the pattern and being extracted from the text data.

In truth finding, one critical thing is to define conflicts: if there was not conflict at all, then a "perfect" model would assume every pattern was reliable and every tuple was true. For the book seller's example, we assume that one book can have only one true author list; so if we knew one list was true, then any different list of the same book would be false. This originated from our commonsense. Fortunately, we also have quite a few commonsense rules for temporal facts, i.e., specific attributes. For example, for *country's president*, we know that

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

For the attribute *sports team's player*, we have commonsense rules:

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

We generalize possible commonsense rules as follows:

- **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.

So, we know that the attribute *country's president* follows C1 and C4; and the attribute *sports team's player* follows C3. Here comes the third challenge (besides O1 and O2): it is necessary to model the commonsense for identifying conflicts, estimating pattern reliability, and finding true temporal facts.

To address the three challenges, we propose a novel Probabilistic Graphical Model with Commonsense Constraints (PGMCC) for finding true temporal facts from the results from pattern-based methods. The given input is the observed frequency of (entity, value, time)-tuples extracted by a particular pattern and attached with a particular type of time signal. We model information source as a pair of pattern and type of time signal. We represent the source reliability as an unobserved variable. It becomes a generative process. We first generate a source. Next we generate a (entity, value, time)-tuple. Then we generate the frequency based on the source reliability and the tuple's trustworthiness (i.e., probability of being a truth). Moreover, we generate variables according to the commonsense rules if needed – the variable counts the values/entities that can be matched to one entity/value with or without a time constraint (at one time) from the set of *true* tuples. Given a huge number of patterns (i.e., 57,472) and tuples (i.e., 116,631) in our experiments, our proposed unsupervised learning model PGMCC can effectively estimate pattern reliability and find true temporal facts.

Our main contributions can be summarized as follows.

- We introduce the idea of PGM-based truth finding to the task of pattern-based temporal fact extraction.
- We propose a new unsupervised probabilistic model with observed constraints to model the reliability of textual patterns, the trustworthiness of temporal tuples, and the commonsense rules for certain types of facts.
- Experimental results show that our model can improve AUC and F1 by more than 7% over the state-of-the-art methods.

The rest of this paper is organized as follows. Section 2 introduces the terminology and defines the problem. Section 3 presents an overview as well as details of the proposed model. Experimental results can be found in Section 4. Section 5 surveys the literature. Section 6 concludes the paper.

2 TEMPORAL TRUTH DISCOVERY

In this section, we first give the terminology and then present the problem definition.

2.1 Terminology

Definition 2.1 (Temporal Fact (Tuple)). Let $\mathcal{F} = \{f_1, f_2, f_3, \ldots\}$ be the set of temporal facts. Each fact f is in the format of (entity, value, time). $\mathcal{F}$ was extracted by textual pattern-based methods [9, 10] and attached with time signals from the text data. The goal is to infer whether the fact tuples are true or false. $\mathcal{F}_{e,v}$ denotes a set of temporal fact f of entity e and value v. $\mathcal{F}_{v,t}$ stands for a set of temporal fact f of value v and time t. Let $\mathcal{L} = \{l_{f_1}, l_{f_2}, l_{f_3}, \ldots\}$ be the set of truth label for $\mathcal{F}$, where $l_{f_i} \in \{0, 1\}$.

Definition 2.2 (Pattern Source). Let $\mathcal{P}^{(*)} = \{p_1^{(*)}, p_2^{(*)}, p_3^{(*)}, \ldots\}$ be the set of pattern source, here $* \in \{``post'', ``tag''\}$ stands for the type of time signal (i.e., "text generation time" and "temporal tag"). One pattern paired with different types of time signals will be treated as different pattern sources.

Definition 2.3 (Extraction). Let $\mathcal{E} = \{e_1, e_2, e_3, \ldots\}$ be the set of extractions. Our generative model will take $\mathcal{E}$ as input. An extraction item e is in the format of (f, $p^{(*)}$, o). Here o stands for the observed frequency of fact tuples f that were extracted by pattern source $p^{(*)}$ in $\mathcal{E}$.

Table 1 presents examples of temporal fact tuples, pattern sources, and extractions (observed frequencies).

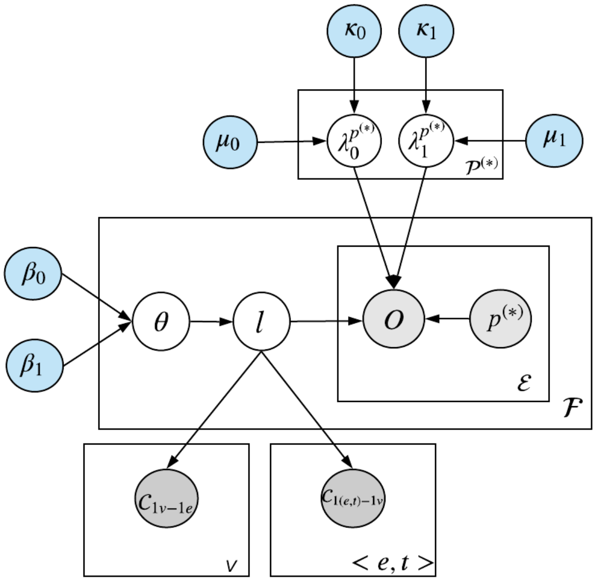

Figure 1: Probabilistic Graphic Model with Commonsense Constraint $\{C_{1(e,t)-1v}, C_{1(v)-1e}\}$

Definition 2.4 (Constraint). Our idea is to represent each commonsense rule as a variable. The variable is likely to be observed as 1. Some examples are as follows:

- one `Value` matches with only one `Entity`, denoted as C_{1v-1e} that counts the number of such entities.
- one `Entity` at one `Time` matches with only one `Value`, denoted as $C_{1(e,t)-1v}$ that counts the number of such values.
- one `Value` at one `Time` matches with only one `Entity`, denoted as $C_{1(v,t)-1e}$ that counts the number of such entities.

Different attributes hold different commonsense constraints. For attribute *country's president*, the constraints are C_{1v-1e} and $C_{1(e,t)-1v}$. For attribute *sports team's player*, the constraint is $C_{1(v,t)-1e}$.

2.2 Problem Definition

Suppose the set of extractions $\mathcal{E}$ have been obtained by pattern-based methods from text data. We define the problem as follows: **Given** a set of extractions $\mathcal{E}$, pattern sources $\mathcal{P}^{(*)}$, and the constraints C_a for attribute a, **infer** truth $\mathcal{T}$ for all temporal facts $\mathcal{F}$ contained in $\mathcal{E}$ and quality information for each pattern source $p^{(*)}$.

3 THE PROPOSED APPROACH

In this section, we first give an overview of our proposed model and then explain the details.

3.1 Overview

As described in the former section, the given input is the observed frequency of (entity, value, time)-tuples extracted by a particular pattern and attached with a particular type of time signal. We model information source as a pair of pattern and type of time signal. We represent the source reliability as an unobserved variable. It becomes a generative process. We first generate a source.

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.

Next we generate a (entity, value, time)-tuple. Then we generate the frequency based on the source reliability and the tuple's trustworthiness. Moreover, we generate variables according to the commonsense constraints. The variables counts the values/entities that can be matched to one entity/value with or without a time constraint (at one time) from the set of true tuples.

3.2 Model details

Figure 1 gives the plate notation of our model. Each node represent a variable. Blue nodes indicate hyper-parameter. Gray nodes stand for observable variable. And white nodes stand for latent variables we want to infer. The concrete meaning of each variable has been given in Table 2. The link from node a to node b means that b is generated from a distribution that takes values of a as parameters. The detailed generative process is as follows.

Temporal fact trustworthiness. For each temporal fact $f \in \mathcal{F}$, we first draw θ_f, i.e., the prior truth probability of fact f, from a *Beta* distribution with hyper-parameter β_0 and β_1:

$$\theta_f \sim Beta(\beta_0, \beta_1). \quad (1)$$

β_0 and β_1 represent the prior distribution of fact reliability. In practice, if we have a strong prior knowledge about how likely all or certain temporal facts are true, we can model it with the corresponding hyper-parameters. Otherwise, if we do not have a strong belief, we set a uniform prior, which means it's equally likely to be true or false, and our model can still infer the truth from other factors. After drawing the θ_f, we generate the truth label l_f from a *Bernoulli* distribution with parameter θ_f:

$$l_f \sim Bernoulli(\theta_f). \quad (2)$$

Pattern source reliability. As aforementioned, a reliable pattern source is more likely to extract true facts with higher counts, and extract false facts with lower counts. Therefore, we choose average count of false/true as latent pattern reliable weight, it's represented as $\lambda_0^{p^{(*)}}, \lambda_1^{p^{(*)}}$ for pattern $p^{(*)}$. The Gamma distribution is utilized because it is the conjugate prior of Poisson distributions. Initially, these two parameters are generated from *Gamma* distribution with hyper-parameter $\{\mu_0, \kappa_0\}/\{\mu_1, \kappa_1\}$, respectively. Here, μ_0 and μ_1 represent the prior number of false/true fact the pattern extract, and κ_0 and κ_1 determine the prior sum of false/true fact count:

$$\lambda_0^{p^{(*)}} \sim Gamma(\mu_0, \kappa_0); \lambda_1^{p^{(*)}} \sim Gamma(\mu_1, \kappa_1) \tag{3}$$

Extraction observation. For each extraction e $\in \mathcal{E}$, it is composed of $\{f, p^{(*)}, o\}$. f_e denotes the temporal fact f belongs to e, $p^{(*)}$ denotes where it's extracted, o_e stands for extraction e's observation count. When the truth label of fact f_e is false, o_e is generated from *Poisson* distribution with $p^{(*)}$'s false speaking side parameter $\lambda_0^{p^{(*)}}$. While f_e is true, o_e is generated from Poisson Distribution with $p^{(*)}$'s true speaking side parameter $\lambda_1^{p^{(*)}}$:

$$\begin{aligned} o_e &\sim Poisson(\lambda_0^{p^{(*)}}) \quad \text{if } l_{f_e} = 0, \\ o_e &\sim Poisson(\lambda_1^{p^{(*)}}) \quad \text{if } l_{f_e} = 1. \end{aligned} \tag{4}$$

Constraints. Finally, we draw the constraint variables. In temporal fact extraction, we define two variables $C_{1(e,t)-1v}$ and C_{1v-1e}. $C_{1(e,t)-1v}$ limits the number of truth on certain constraint key $\{e, t\}$. There are as many $C_{1(e,t)-1v}$ variables as unique $\{e, t\}$ keys:

$$C_{e,t} = \sum_f l_f, \quad f \in \mathcal{F}_{e,t}, \tag{5}$$

where $\mathcal{F}_{e,t}$ denotes a set of f with same $\{e, t\}$. Each $C_{1(e,t)-1v}$ is generated by $\mathcal{F}_{e,t}$ set.

C_{1v-1e} ilimits the truth of fact with same $\{v\}$:

$$C_v = \sum_{e \in E} l_{e,v} \quad \begin{cases} l_{e,v} = 1, & \text{if} \quad \exists l_f = 1, \quad f \in \mathcal{F}_{e,v}; \\ l_{e,v} = 0, & \text{otherwise.} \end{cases} \tag{6}$$

where $\mathcal{F}_v$ denotes set of fact with value v, $\mathcal{F}_{e,v}$ stands for a set of temporal fact f with same $\{e, v\}$, $\mathcal{F}_{e,v} \in \mathcal{F}_v$. $l_{e,v}$ denoted the truth label of v, e. Each C_{1v-1e} is generated by $\mathcal{F}_v$. If there is true fact f $\in \mathcal{F}_{e,v}$, then $l_{e,v}$ equals to 1, otherwise, $l_{e,v}$ equal with 0.

3.3 Inference Algorithm

In this section, we explain the approach to estimate the latent variables in our probabilistic model. We use the Markov Chain Monte Carlo (MCMC) methods to approximate the fact truth label and pattern reliability weight. It is an iterative sampling method to infer the truth label for each fact with observable variable and current truth labels of other facts.

We first sampled l_f for each $f \in \mathcal{F}_{e,t}$

$$\begin{aligned} &p(l_f = i | \mathcal{L}^{-f}, \mathcal{O}, \mathcal{P}^{(*)}) \\ &\propto p(l_f = i | \mathcal{L}^{-f}) \times \prod_{e \in \mathcal{E}_f} p(o_e, p_e^{(*)} | l_f = i, o^{-f}, p^{(*)^{-f}}), \end{aligned} \tag{7}$$

where $\mathcal{L}^{-f}$ denotes the truth label of all facts in $\mathcal{F}$ except f, $p^{(*)^{-f}}$ stands for the pattern source $p^{(*)}$ expecting fact f, and o^{-f} denotes that the set of observations o except those of with fact f. The first part of Eq.(7) can be rewritten as:

$$p(l_f = i | \mathcal{L}^{-f}) = \frac{\beta_i}{\beta_1 + \beta_0}. \tag{8}$$

Because Gamma distribution is the conjugate prior of *Poisson* distributions, each item in Eq.(7) can be rewritten as:

$$p(o_e, p_e^{(*)} | l_f = i, o^{-f}, p^{(*)^{-f}}) \propto Poisson\left(o_e | \frac{n^{-f}_{p_e^{(*)}, sum, i} + \kappa_i}{n^{-f}_{p_e^{(*)}, num, i} + \mu_i} \right), \tag{9}$$

where $n^{-f}_{p_e^{(*)}, sum, i}$ denotes the sum of extraction e's observation count o_e come from $p_{(*)}$. Here the referred fact is not f and its truth label is i. $n^{-f}_{p_e^{(*)}, num, i}$ denotes the number of extraction e come from $p^{(*)}$ where the fact is not f and its truth is i.

Hence by omitting constant $\beta_1 + \beta_0$, Eq.(7) can be rewritten as:

$$\begin{aligned} p(l_f = i | \mathcal{L}^{-f}, \mathcal{O}, \mathcal{P}^{(*)}) &\propto \beta_i \prod_{e \in \mathcal{E}_f} \frac{e^{-\lambda} \lambda^{o_e}}{o_e!}, \\ \text{where,} \quad \lambda &= \frac{n^{-f}_{p_e^{(*)}, sum, i} + \kappa_i}{n^{-f}_{p_e^{(*)}, num, i} + \mu_i}. \end{aligned} \tag{10}$$

After sampling l_f for each $f \in \mathcal{F}_{e,t}$, we maximize the posterior with constraint. Since the summation of $p(t_f = 1)$ and $p(t_f = 0)$ equals to 1, we normalize $p(t_f)$ for all $f \in \mathcal{F}_{e,t}$ and select the fact f with highest probability as true. Then we update truth probability of $f \in \mathcal{F}_{e,t}$ before we sample the next $\mathcal{F}_{e,t}$ set:

$$\begin{aligned} &p(l_f = 1 | \mathcal{L}^{-f}, \mathcal{O}, \mathcal{P}^{(*)}, C_{1(e,t)-1v}) \\ &= \begin{cases} p(l_f), & \text{if } p(l_f) = MAX\{p(l_{f_{e,t}} = 1), f_{e,t} \in \mathcal{F}_{e,t}\}; \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{11}$$

where $f \in \mathcal{F}_{e,t}$, $\{p(l_{f_{e,t}} = 1), f_{e,t} \in \mathcal{F}_{e,t}\}$ denotes a set of $p(l_{f_{e,t}} = 1)$ probability, and $p(l_f)$ represent its normalized result for Eq.(10).

After sampling $\mathcal{L}$ under constraint $C_{1(e,t)-1v}$, we sample the $l_{e,v}$ with constraint C_{1v-1e} in order to constrain the truth of $\mathcal{L}$:

$$p(r_{e,v} = 1 | \mathcal{T}_v) = \frac{\sum_{f \in \mathcal{F}_{e,v}} l_f}{\sum_{f' \in \mathcal{F}_v} l_{f'}}. \tag{12}$$

$$\begin{aligned} &p(r_{e,v} = 1 | \mathcal{L}_v, C_{1v-1e}) \\ &= \begin{cases} p(r_{e,v}), & \text{if } p(r_{e,v}) = MAX\{p(r_{e,v} = 1), e \in \mathcal{E}\}; \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{13}$$

where Eq.(12) denotes the true probability for value v combined with entity e, and $r_{e,v}$ stands for the truth label of $\{e, v\}$. $\mathcal{L}_v$ represents a set of truth label fact with value v.

Then, we integrate the truth probability of fact f with C_{1v-1e}:

$$\begin{aligned} &p(l_f = 1 | \mathcal{L}^{-f}, \mathcal{O}, \mathcal{P}^{(*)}, C_{1v-1e}, C_{1(e,t)-1v}) \\ &= \begin{cases} p(l_f = 1 | \mathcal{L}^{-f}, \mathcal{O}, \mathcal{P}^{(*)}, C_{1(e,t)-1v}), & \text{if} \quad r_{e,v} = 1; \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{14}$$

Through this inference algorithm, during every iteration, the constrained variables will ensure that there is no commonsense conflict in inferred truth. So it avoids erroneous estimation about pattern source reliability and temporal fact truth.

4 EXPERIMENTS

In this section, we first describe the dataset and then present experimental results.

Method	**Constraints**		**Evaluation Setting**			
	e: Country; v: Person; t: year		On (e,v,t)		On $(e,v,[t_{min},t_{max}])$	
	$C_{1(v)-1e}$	$C_{1(e,t)-1v}$	**AUC**	**F1**	**AUC**	**F1**
TruthFinder [27]	✗	✗	0.0006	0.0012	0.0006	0.0012
LTM [31]	✗	✗	0.1319	0.0199	0.2030	0.0218
LTM [31]	✗	✔	0.0212	0.0505	0.0407	0.0793
TruePIE [10]	✔	✗	0.0587	0.1430	0.0587	0.1430
MajVote [6]	✗	✔	0.3336	0.4318	0.4958	0.5927
TFWIN [24]	✗	✔	0.4746	0.6361	0.5523	0.6489
TFWIN [24]	✔	✔	0.4746	0.6361	0.5523	0.6489
Ours (PGMCC)	✗	✔	0.4840	0.6502	0.6006	0.7254
Ours (PGMCC)	✔	✔	**0.4987**	**0.6634**	**0.6075**	**0.7316**

Table 3: Our proposed model performs better than baseline methods on finding temporal facts.

4.1 Dataset

It has 9,876,086 news articles (4 billion words) published from 1994–2010. We focus on attribute *country's president*. We have 57,472 patterns, 116,631 temporal fact tuples, and 1,326,164 extractions. We collected ground truth from Google and Wikipedia. It includes 3,175 true temporal facts of 130 countries.

4.2 Experiment Settings

4.2.1 Competitive methods. We compare our model with:
- TruthFinder [27]: It was a bootstrapping algorithm for structured data using C_{1v-1e}.
- LTM[31]: It was a probabilistic model, assuming that the truth about an object contains more than one value. We set "object" as {entity, time} and set value as the temporal fact's value.
- TruePIE [10]: It was a bootstrapping method using $C_{1(v)-1e}$ and estimating pattern reliability for fact extraction.
- MajVote [6]: It used the weighted majority voting strategy and returned the most frequent temporal fact.
- TFWIN [24]: It was the state-of-the-art bootstrapping method for truth discovery on fact extraction. However, error propagation is serious in its iterative process. It could not estimate pattern reliability with the general data distributions.

4.2.2 Evaluation settings. All the methods can only find truth of temporal fact at one time point, e.g., (French, Jacques Chirac, 1995). However, due to the incompleteness of fact description in data, some time points of temporal facts could be missing. One way to improve the evaluation is to composite true temporal fact time points $\{e, v, t\}$ into temporal fact time period $\{e, v, [t_{min}, t_{max}]\}$.

We evaluate the performance on both temporal fact time point $\{e, v, t\}$ and temporal fact time period $\{e, v, [t_{min}, t_{max}]\}$. To evaluate on time period $\{e, v, [t_{min}, t_{max}]\}$, we look at every single time points (e, v, t) in the period $(t \in [t_{min}, t_{max}])$.

4.2.3 Evaluation metrics. We evaluate our proposed model and all competitive baselines using standard Information Retrieval metrics: *precision*, *recall*, *F1 score*, and *AUC* (Area Under the Curve). Precision is the the fraction of temporal fact truth among all the temporal fact that were labelled as true. Recall is the fraction of true temporal facts our approach finds among the ground truth temporal facts. F1 score is the harmonic mean of precision and recall. For all of the metrics, higher score indicates that the method has better performance.

4.3 Effectiveness

The results are given in Table 3. Our proposed method PGMCC consistently outperforms all the baselines on finding (country, president, time)-facts (i.e., presidential terms). We carefully compare our method with each baseline and analyze the comparison results.

PGMCC vs MajVote: Table 3 shows that PGMCC performs significantly better than MajVote (+16.6% AUC; +23.3% F1) on evaluating time points and performs better with (+11.2% AUC; +14.6% F1) on evaluating time periods. The reason is that MajVote simply returned the frequent facts and could not return correct value while the true facts were minority. PGMCC estimates pattern source reliability and can infer the trustworthiness of fact tuples based on the reliable pattern sources.

PGMCC vs LTM: PGMCC performs significantly better than LTM (+34.5% AUC; +64.4% F1) on evaluating time points, and performs better with (+40.45% AUC; +71.2% F1) on evaluating time periods. LTM was designed to solve structured truth finding like the bookseller example. So, there were many conflicts when applied to temporal fact extraction. PGMCC has multi-constraint as observable variables to alleviate the issue.

PGMCC vs TFWIN: PGMCC performs better than TFWIN (+2.4% AUC; +2.8% F1) on evaluating time points, and performs better with (+5.2% AUC; +8.3% F1) on evaluating time periods. TFWIN started with seed patterns and defined constraints as a rule to eliminate conflicting tuples. However, the inference on conflicts was based on local information (i.e., the current pattern reliability estimation). During this process, error might propagate through iterations. PGMCC is a probabilistic graphical model that can avoid error propagation by modeling constraints as variables and inferring truth with the global data distributions.

PGMCC with different constraints: The last two rows in Table 3 shows the results. For both PGMCC and TFWIN models, a complete constraint set, i.e., $\{C_{1(v)-1e}$ and $C_{1(e,t)-1v}\}$, gives the best performance. Partial constraint cannot fully identify conflicts or false tuples. In the task of extracting *country's president*, $C_{1(e,t)-1v}$ plays a significant role.

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.

4.4 Case Studies

4.4.1 False case analysis. Table 4 presents the results. Without using the constraint C_{1v-1e}, the model predicted *j.r_chirac* served four countries (France, Spain, Greece and Tunisia) as president. This is never truth. PGMCC of complete constraints $C_{1(e,t)-1(v)}$ and $C_{1v-1(e)}$ can solve the problem.

4.4.2 Pattern source reliability analysis. Our model gives two reliability scores for each pattern - one is on the probability of extracting true tuples $\lambda_1^{p^{(*)}}$; and the other is the probability of extracting false tuples $\lambda_0^{p^{(*)}}$. Note that they are not on the probability of the pattern’s extractions being true or false. So the sum is not 1. We define a scoring function to summarize the two scores:

$$r_{p^{(*)}} = \frac{\lambda_1^{p^{(*)}}}{\lambda_1^{p^{(*)}} + \lambda_0^{p^{(*)}}}. \tag{15}$$

Table 5 presented some pattern examples and their scores. Here are our observations. First, the pattern “president Person of Country” is the only pattern that shows high reliability on both types of time signals (above 0.85). Second, the textual patterns that describe the current presidency are likely to have higher reliability on *text generation time* (“post”) than *temporal tag* (“tag”), because the presidency was likely to be in the same time as the document was generated. These patterns usually have words such as “current”, “newly”, and “now”. Third, the textual patterns that describe the past presidency are likely to have higher reliability on *temporal tag* (“tag”) than *text generation time* (“post”), because the presidency was likely to be in the same time as the event (described in the sentence) happened but before the time of the document being generated. These patterns usually have words such as “have governed”, “have ruled”, “former”, and “formerly”. Forth, quite a few patterns have reliability scores that are not very high nor very low. In the examples, the patterns have words ”ruled” and “signed”, however, it is hard to tell the time was long before or just a few hours/days before the document was generated. It could be about a former president. It could be about the current president, too. Lastly, we found that the textual patterns off the topic would have very low reliability. For example, when the patterns have words “premier”, “foreign minister”, “golfer”, and “basketball player”, their reliability scores on both types of time signals were close to zero.

Textual Pattern p	$r_{p(post)}$	$r_{p(tag)}$
president Person of Country	0.920	0.870
Country’s current president Person,	0.978	0.250
Country’s newly elected president , Person ,	0.970	0.030
Person, now president of Country,	0.750	0.110
Person, who has ruled Country	0.438	0.994
\$COUNTRY’s former president Person	0.113	0.994
Person, who ruled Country	0.607	0.758
Country president Person signed	0.553	0.327
Country premier Person	0.012	0.010
Country foreign minister Person	0	0
Country golfer Person	0	0

Table 5: Pattern’s reliability scores for country’s presidency.

5 RELATED WORK

In this section, we review two relevant fields to our work, temporal fact extraction and truth discovery.

5.1 Truth Discovery

In big data era, the issue of “Veracity” on resolving conflicts among multi-source information is quite serious [1, 4, 5, 13, 14, 22, 23, 25, 28]. Truth discovery methods find trustworthy information from conflicting multi-source [11, 12, 15, 16, 26]. Several truth discovery methods have been proposed for various scenarios, and they have been successfully applied in diverse application domains. A few truth discovery methods are probabilistic model. LTM solved the “Book’s author list problem” and modeled its source in two-fold quality [32]. GTM solved the task of finding true numeric value of “New York City’s population” [31]. TEXTTRUTH found the true answer for a question from multi users [30].

5.2 Temporal Fact Extraction

Temporal fact extraction is to extract (entity, attribute name, attribute value)-tuples along with their time conditions from text corpora [2, 8, 20, 21, 29]. Textual patterns have been proposed to extract structured data from unstructured text data in an unsupervised way, such as E-A patterns [7], parsing patterns [17], and meta patterns [9]. However, patterns are of different reliability and extractions are sometimes conflicting. In order to get reliable temporal fact, we addressed this problem using truth discovery.

6 CONCLUSIONS

In this work, we proposed a probabilistic graphical model for inferring true facts and pattern reliability. It had two novel designs for temporal facts: (1) it modeled pattern reliability on temporal tag in text and text generation time; (2) it modeled commonsense constraints as observable variables. Experimental results demonstrated that our model outperformed existing methods.

REFERENCES

[1] Laure Berti-Equille. 2015. Data veracity estimation with ensembling truth discovery methods. In *Big Data (Big Data), 2015 IEEE International Conference on*. IEEE, 2628–2636.

[2] Melisachew Wudage Chekol. 2017. Scaling probabilistic temporal query evaluation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 697–706.

[3] Luciano Del Corro, Abdalghani Abujabal, Rainer Gemulla, and Gerhard Weikum. 2015. Finet: Context-aware fine-grained named entity typing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 868–878.

[4] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. 2009. Integrating conflicting data: the role of source dependence. *Proceedings of the VLDB Endowment* 2, 1 (2009), 550–561.

[5] Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. 2010. Corroborating information from disagreeing views. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 131–140.

[6] Sally A Goldman and Manfred K Warmuth. 1995. Learning binary relations using weighted majority voting. *Machine Learning* 20, 3 (1995), 245–271.

[7] Rahul Gupta, Alon Halevy, Xuezhi Wang, Steven Euijong Whang, and Fei Wu. 2014. Biperpedia: An ontology for search applications. *Proceedings of the VLDB Endowment* 7, 7 (2014), 505–516.

[8] Tuan-Anh Hoang-Vu, Huy T Vo, and Juliana Freire. 2016. A unified index for spatio-temporal keyword queries. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 135–144.

[9] Meng Jiang, Jingbo Shang, Taylor Cassidy, Xiang Ren, Lance M Kaplan, Timothy P Hanratty, and Jiawei Han. 2017. Metapad: Meta pattern discovery from massive text corpora. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 877–886.

[10] Qi Li, Meng Jiang, Xikun Zhang, Meng Qu, Timothy P Hanratty, Jing Gao, and Jiawei Han. 2018. Truepie: Discovering reliable patterns in pattern-based information extraction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 1675–1684.

[11] Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. 2014. A confidence-aware approach for truth discovery on long-tail data. *Proceedings of the VLDB Endowment* 8, 4 (2014), 425–436.

[12] Qi Li, Yaliang Li, Jing Gao, Bo Zhao, Wei Fan, and Jiawei Han. 2014. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. ACM, 1187–1198.

[13] Xian Li, Weiyi Meng, and T Yu Clement. 2016. Verification of Fact Statements with Multiple Truthful Alternatives.. In *WEBIST (2)*. 87–97.

[14] Xian Li, Weiyi Meng, and Clement Yu. 2011. T-verifier: Verifying truthfulness of fact statements. In *Data Engineering (ICDE), 2011 IEEE 27th International Conference on*. IEEE, 63–74.

[15] Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2016. A survey on truth discovery. *ACM Sigkdd Explorations Newsletter* 17, 2 (2016), 1–16.

[16] Yaliang Li, Qi Li, Jing Gao, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2015. On the discovery of evolving truth. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 675–684.

[17] Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. 2012. PATTY: a taxonomy of relational patterns with semantic types. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 1135–1145.

[18] Nils Reimers, Nazanin Dehghani, and Iryna Gurevych. 2016. Temporal anchoring of events for the timebank corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 2195–2204.

[19] Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R Voss, Heng Ji, Tarek F Abdelzaher, and Jiawei Han. 2017. Cotype: Joint extraction of typed entities and relations with knowledge bases. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1015–1024.

[20] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering* 30, 10 (2018), 1825–1837.

[21] Avirup Sil and Silviu-Petru Cucerzan. 2014. Towards Temporal Scoping of Relational Facts based on Wikipedia Data. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. 109–118.

[22] VG Vydiswaran, ChengXiang Zhai, and Dan Roth. 2011. Content-driven trust propagation framework. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 974–982.

[23] Dalia Attia Waguih and Laure Berti-Equille. 2014. Truth discovery algorithms: An experimental evaluation. *arXiv:1409.6428* (2014).

[24] Xueying Wang, Haiqiao Zhang, Qi Li, Yiyu Shi, and Meng Jiang. 2019. A Novel Unsupervised Approach for Precise Temporal Slot Filling from Incomplete and Noisy Temporal Contexts. In *The World Wide Web Conference*. ACM, 3328–3334.

[25] Houping Xiao, Jing Gao, Qi Li, Fenglong Ma, Lu Su, Yunlong Feng, and Aidong Zhang. 2016. Towards confidence in the truth: A bootstrapping based truth discovery approach. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1935–1944.

[26] Houping Xiao, Yaliang Li, Jing Gao, Fei Wang, Liang Ge, Wei Fan, Long H Vu, and Deepak S Turaga. 2015. Believe it today or tomorrow? detecting untrustworthy information from dynamic multi-source data. In *Proceedings of the 2015 SIAM International Conference on Data Mining*. SIAM, 397–405.

[27] Xiaoxin Yin, Jiawei Han, and S Yu Philip. 2008. Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering* 20, 6 (2008), 796–808.

[28] Xiaoxin Yin and Wenzhao Tan. 2011. Semi-supervised truth discovery. In *Proceedings of the 20th international conference on World wide web*. ACM, 217–226.

[29] Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian Sadler, Michelle Vanni, and Jiawei Han. 2018. TaxoGen: Constructing Topical Concept Taxonomy by Adaptive Term Embedding and Clustering. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

[30] Hengtong Zhang, Yaliang Li, Fenglong Ma, Jing Gao, and Lu Su. 2018. Texttruth: an unsupervised approach to discover trustworthy information from multi-sourced text data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2729–2737.

[31] Bo Zhao and Jiawei Han. 2012. A probabilistic model for estimating real-valued truth from conflicting sources. *Proc. of QDB* (2012).

[32] Bo Zhao, Benjamin IP Rubinstein, Jim Gemmell, and Jiawei Han. 2012. A bayesian approach to discovering truth from conflicting sources for data integration. *Proceedings of the VLDB Endowment* 5, 6 (2012), 550–561.