Non-Named Entities and Vagueness

Pierre-Henri PARIS, Fabian SUCHANEK

The NoRDF project

- **Extract and model complex information from natural language text:**
  - **events,**
  - **causation,**
  - **conditions,**
  - **precedence,**
  - **stories,**
  - **negation,**
  - **beliefs,**
  - **sentiment.**

- **https://nordf.telecom-paris.fr/en/**

Outline

1. **Non-Named Entities – The Silent Majority**

2. **The Vagueness of Vagueness in Noun Phrases**

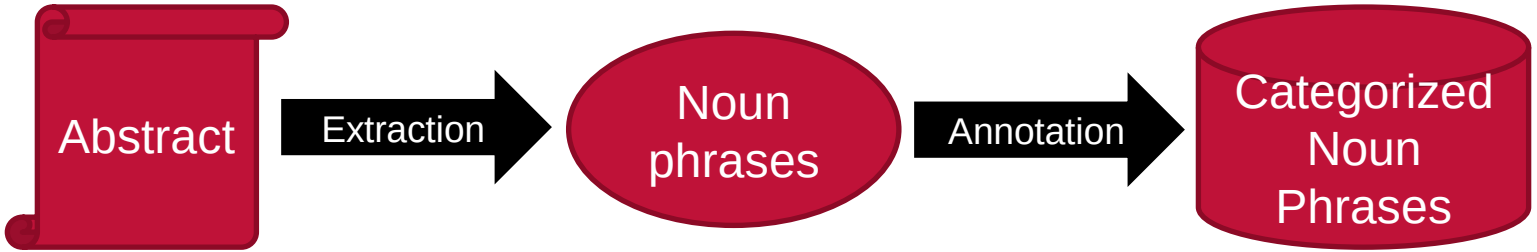
Noun phrases

"*The Arab Spring* **resulted in** *contentious battle between a consolidation of power by religious elites* **and** *the growing support for democracy*."

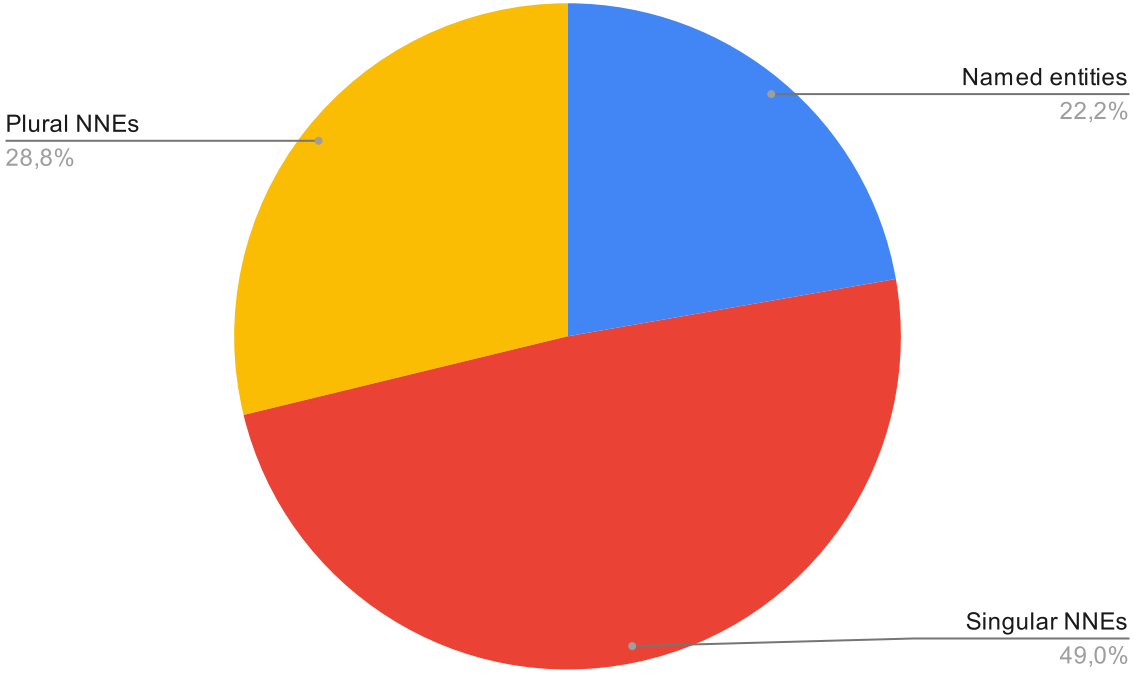
- **This is a factual sentence, and thus interesting for information extraction.**
- **However, it contains only a single named entity. Many fact-extraction techniques will simply not consider the other non-named entities.**

Case Study: Wikipedia

- **Wikipedia featured articles (high quality articles)**
- **Wikipedia is a widely used standard reference, in both research and industry applications**
- **1 article abstract for each of the 30 featured article topics**

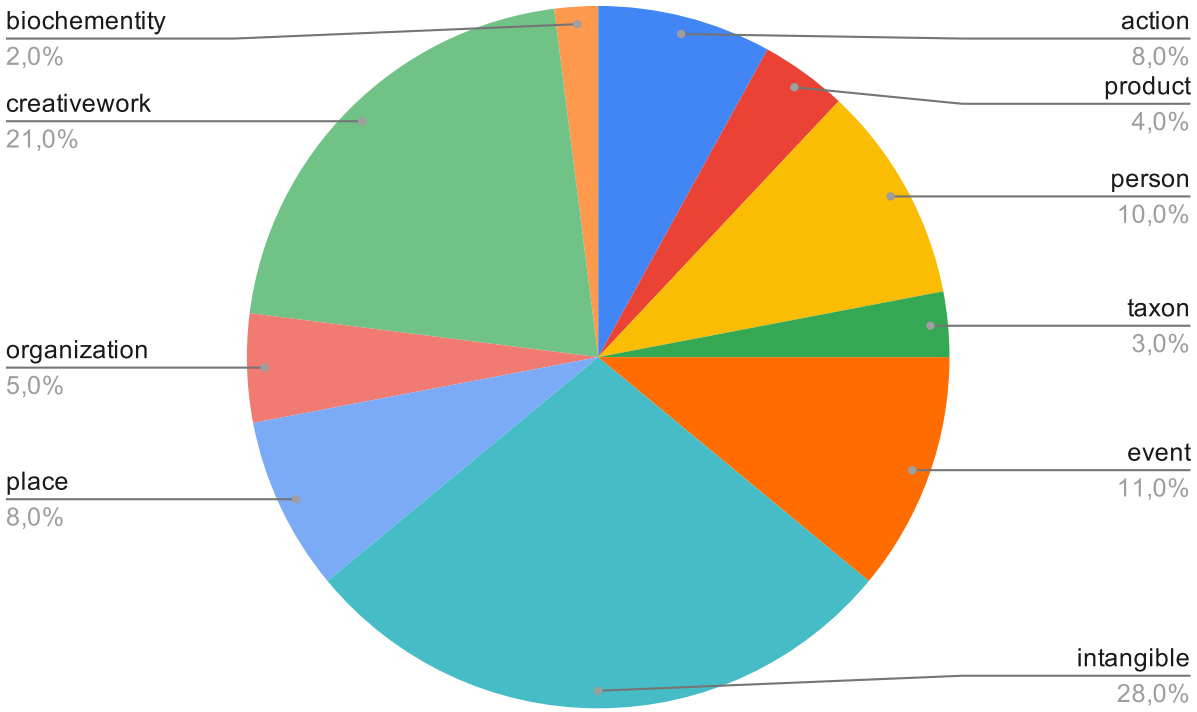


Manual Study of Non-Named Entities in Wikipedia

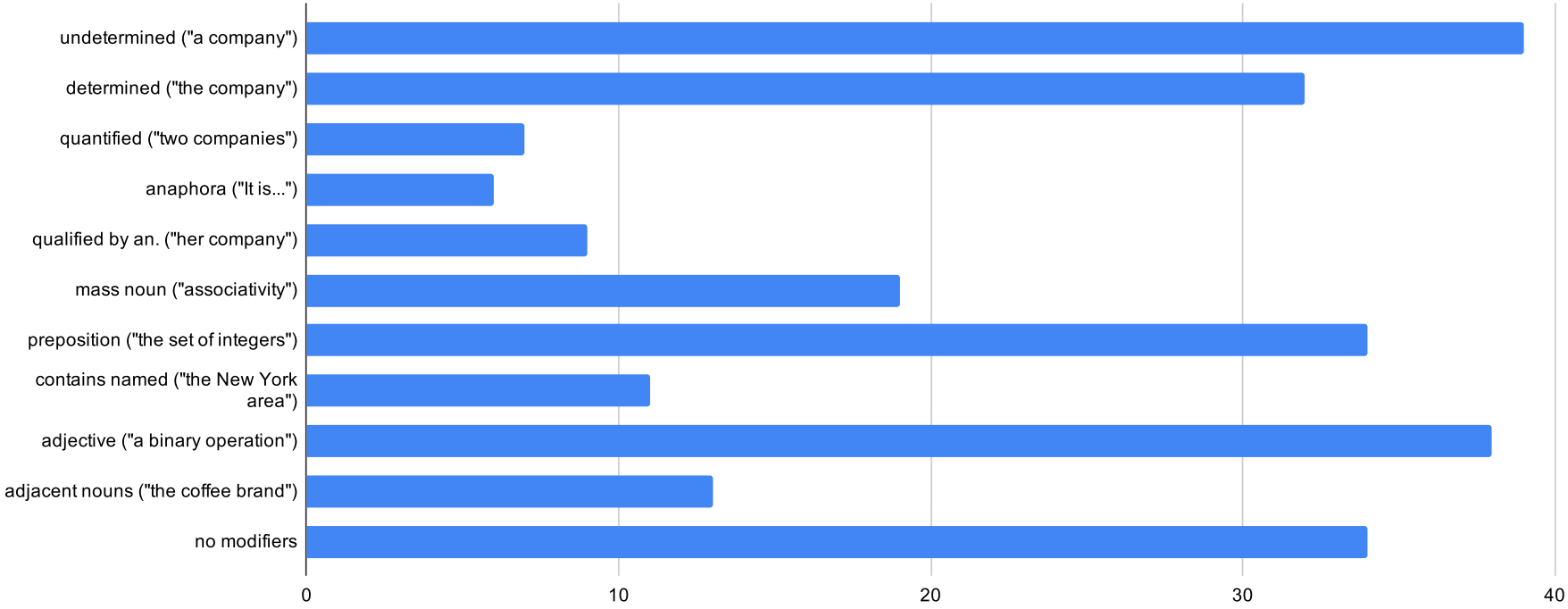


1 article abstract for each of the 30 featured Wikipedia article categories

Non-named entities by Yago class

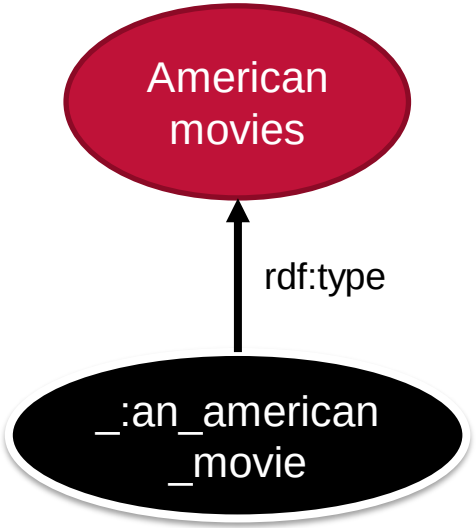


Non-named entities by nature and modifiers



Adding Non-named Entities to RDFS Knowledge Bases

- **Extracting non-named entities**
  - Possible with Semantic parsers, AMR, Open Information Extraction systems…
  - Named entities are mapped to the knowledge base
- **Modeling non-named entities**
  - Replace anaphoras and determined noun phrases with their referent
  - Simple nested noun phrases are linked together
  - Use of ad-hoc classes for plural noun phrases
  - Make singular noun phrases anonymous instances of ad-hoc classes
  - Replication of numbered noun phrases
  - Mass nouns became classes



The Gap that remains

- **Knowledge representation:**
  - **Some knowledge bases contain mainly classes, others mix classes and instances, and others duplicate them**
  - **Plural noun phrases like « *hundreds of soldiers* »**
  - **Vague noun phrases like « *large-scale settlement* »**
  - **Nested entities like « *the growing support for democracy in many Muslim-majority states* »**

- **Canonicalization: Similar non-named entities in different contexts like two different rises of the same stock market**

- **Facts: Comparatives, superlatives, and temporal comparisons like « *A rose is more beautiful than a daisy* »**

Outline

1. **Non-Named Entities – The Silent Majority**

2. **The Vagueness of Vagueness in Noun Phrases**

The Vagueness Problems

"In the early 20th century, German researchers made advances in linking smoking to health harms."

- **This is a factual sentence, and thus interesting for information extraction.**

  However, when is the early 20th century? Who are the German researchers? How many of them? What advances? What are the health harms?

- **How to detect, extract, model and reason with vagueness?**

- **Vagueness has not been sufficiently studied**

Types of Vagueness

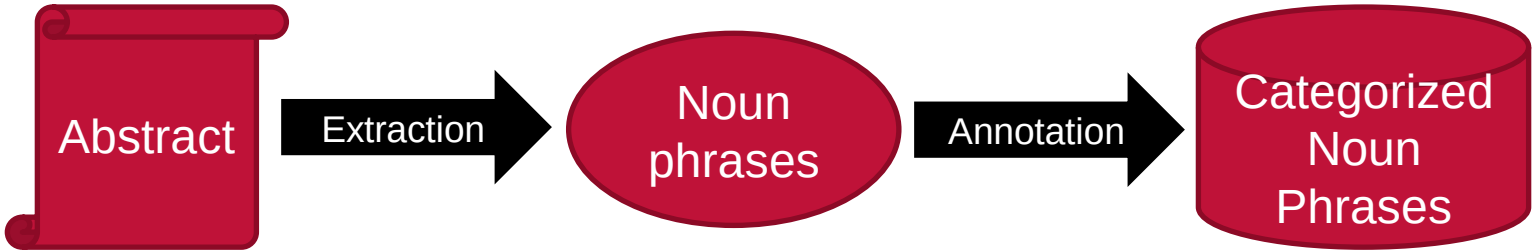
- **Scalar vagueness**
  - **expressions that can be interpreted as a scalar**
  - *"a rich person", "long-running debate", "early history"*
- **Quantitative vagueness**
  - **Expressions that refer to an unspecified portion of an entity or to a set of entities whose number is not identified**
  - *"a part of the film", "many scientists", "scientists"*
- **Subjective vagueness**
  - **Expressions that can apply to a certain degree, and where there is no consensus on how to measure this degree**
  - *"a beautiful man", "high ethical standards", "tensions"*

Vagueness often depends on its context, i.e. the sentence in which it is used:
"Computer scientists believe that AES encryption is safe" **VS** *"Computer scientists did not exist in the Middle Ages"*

Case Study: Wikipedia

- Wikipedia featured articles (high quality articles)
- Wikipedia is a widely used standard reference, in both research and industry applications
- 1 article abstract for each of the 30 featured article topics

⟹ our study constitutes a probable lower bound on the frequency of vagueness



Case Study: Wikipedia

Vagueness	Number	Proportion	Examples
None	1816	74.1%	“the company”, “three children”, ...
Scalar	142	5.8%	“long-running debate”, “early history”, “a large share”, ...
Quantitative	339	13.8%	“many of the ideas”, “government bodies”, “other media”, ...
Subjective	235	9.6%	“tensions”, “high ethical standards”, “sufficient interest”, ...

⇒ ¼ of noun phrases are vague!

Existing Work: Detecting Vagueness

- **Noun phrase level:**
  - **Vague definitions in ontologies:**
    certain words indicate vagueness
  - **Requirement specifications in the railway domain:**
    difficult to define phrases that identify vagueness with a high accuracy
  - **Privacy policies:**
    collection of vague words to train a classifier, there is no exhaustive list of vague words
- **Sentences and noun phrases levels:**
  - **Crowdsourcing annotations in privacy policies to train two classifiers:**
    vagueness is context-dependent

⟹ A Ridge classifier achieved an accuracy of 0.78 on our corpus

Existing Works: Reasoning with Vagueness

- **Logical Approaches**
  - Generalized quantifiers for quantitative vagueness
  - Fuzzy Logic for scalar vagueness
  - Fuzzy Logic + $\mathcal{ALC}$ Description Logic for scalar vagueness

- **Probabilistic Approaches**
  - Bayesian approach for scalar vagueness in adjectives
  - Word co-occurrence for scalar vagueness in modifiers

⟹ Quantitative and scalar vagueness could be addressed if different approaches are used together, but subjective vagueness remains out of reach

Conclusion

- **Vagueness is a common phenomenon in noun phrases, but current approaches do not address it comprehensively**
- **The detection and categorization of vagueness can be performed**
- **Modeling vagueness is more complex:**
  - **Many approaches exists for scalar vagueness**
  - **Quantitative vagueness could be modeled with Generalized Quantifiers or with a combination of anonymous instances and axioms or with ad-hoc classes**
  - **Subjective vagueness could be annotated with the person who employed it**

Additional material

Case Study: Wikipedia – Vagueness in General

- 26% of noun phrases

- 38% of intangible objects
- 35% of Taxon noun phrases
- 35% of all actions
- The classes of person, place, organization, and product are rarely vague

- 46% of undetermined noun phrases
- 45% of noun phrases with an adjective
- 13% of determined noun phrases

Case Study: Wikipedia – Scalar Vagueness

- The least frequent vagueness type (6%)
- 68% of the scalar vague noun phrases is induced by an adjective
- Scalar vague noun phrases often relate to time and space
  - 25% of the scalar vague noun phrases are events
- 32% The main other class of scalar vagueness is intangible objects

Case Study: Wikipedia – Quantitative Vagueness

- The most frequent type of vagueness with 14% of the noun phrases

- 44% of plural noun phrases are quantitatively vague
- 95% of quantitative vagueness concerns plural noun phrases
- 67% of quantitative vagueness concerns undetermined nouns
- 19% of quantitatively vague noun phrases come with a vague quantity, such as “several”, “many”, or “few”

- Only 6% of quantitative vagueness concerns

- 26% of quantitative vagueness concerns creative works

Case Study: Wikipedia – Subjective Vagueness

- 10% of the noun phrases are subjectively vague

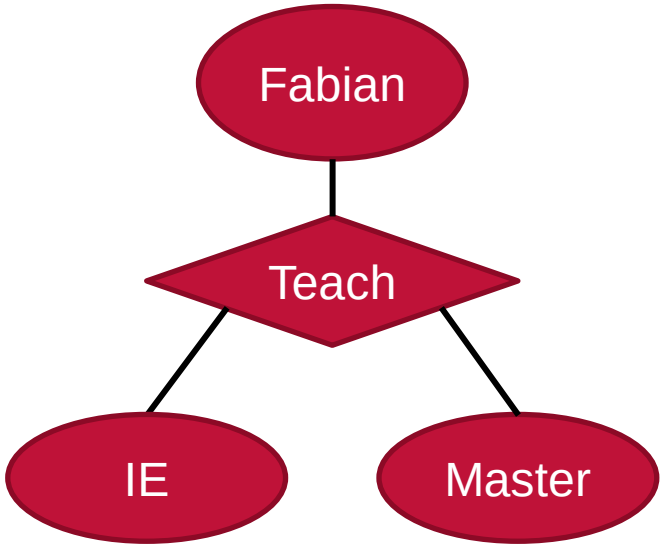
- 52% of subjective vagueness concerns intangible objects
- 16% of the creative works are subjectively vague

- 61% of subjective vagueness comes in noun phrases with adjectives
- 32% of subjective vagueness concerns mass nouns

NoRDF

What we want to model: High arity relationship

- **Arity is the number of arguments of a relationship**
- **Because information is not always binary, we want to model relationship with arity > 2**
- ***"Fabian teaches Information Extraction in a specialized Master's program."***
- **Reification and other constructs that tackle arbitrary relationships lead to undecidability of reasoning, if used in an unrestricted way.**



What we want to model: Events

Events are objects in time

Many modeling/reasoning dimensions depend on events

What, when, topic, where and who?

> **"*The Apollo 11 crew landed on the Moon on July 21, 1969.*"**

Evolution of long-lasting event? Use of a temporal logic or Event calculus?

What we want to model: Precedence

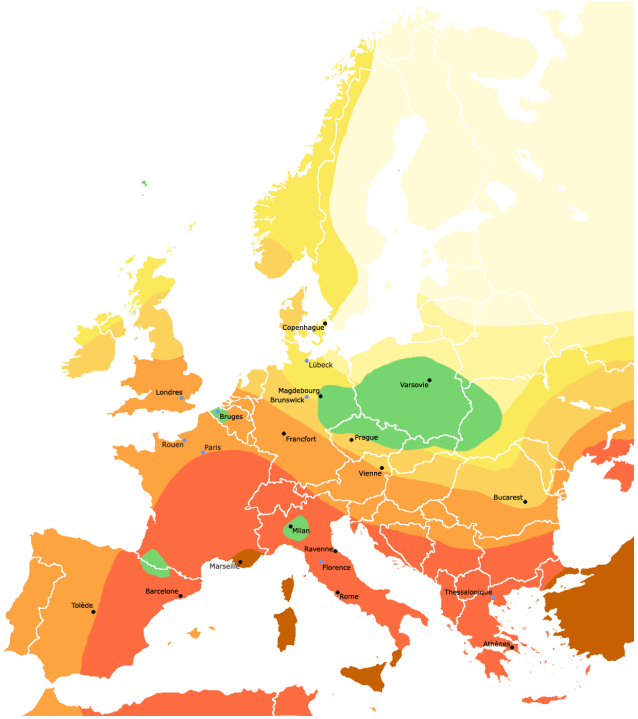
An event that happened before another one

Useful to model a narrative

$\forall E_i : before(E_1, E_2), before(E_2, E_3) \Rightarrow before(E_1, E_3)$

"WW1 happened before WW2 and WW2 happened before the creation of European Union" ⟹ "WW1 happened before the creation of European Union"

How to model temporal dimension? Point-based or interval?



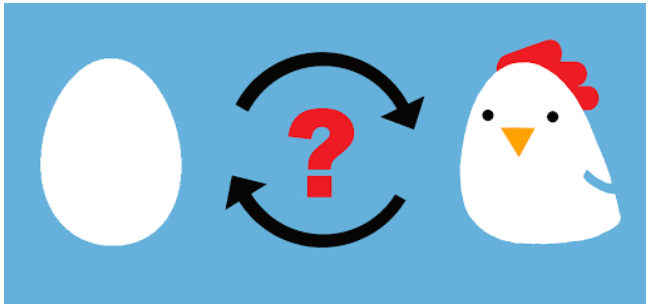
What we want to model: Causation

An event contributes to the production of another event

Causation is necessary to explain succession of events

$\forall \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$: $causes(a, b), causes(b, c) \Longrightarrow causes(a, c)$

"The crisis of 2008 caused a recession. The number of homeless families is increasing due to the recession."



Multiple implementations are possible

- **Necessary causes**
  - $\forall \boldsymbol{x}, \boldsymbol{y}$: $necessary(x, y)\ and\ y \implies x$
  - "*People who have a salary necessarily have a job.*"
- **Sufficient causes**
  - $\forall \boldsymbol{x}, \boldsymbol{y}$: $sufficient(x, y)\ and\ x \implies y$
  - "*The recession caused the increase in the number of homeless families.*"
- **Contributory causes**
  - $\forall \boldsymbol{x_i}, \boldsymbol{y}$: $contributory(\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}, y), \boldsymbol{x_1}, \ldots, \boldsymbol{x_n} \implies y$
  - "*Excessive risk-taking by banks contributed to the crisis of 2008.*"

What we want to model: Negation

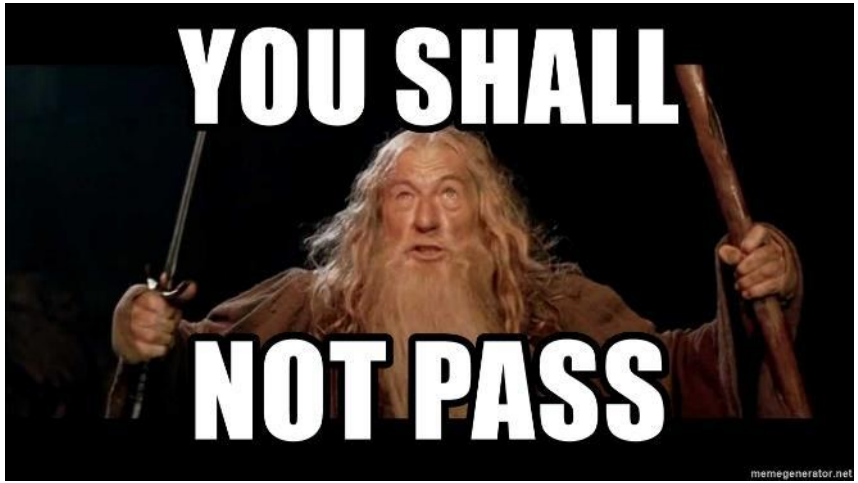
Modeling of false facts

Most knowledge bases contain only positive facts, yet many interesting facts are negative

$\forall x\colon deathdate(x, d) \Rightarrow \neg alive(x)$

"Elvis died in 1977" ⟹ "Elvis is not alive"

The complexity increases as the arity of the negated relationship increases



What we want to model: Beliefs

A belief is an attitude that some proposition about the world is true

Allows to have contradicting statements in the same KB

> **"*Flat-earthers believe the earth is flat, but the earth is almost a sphere.*"**

$\forall \boldsymbol{\phi}: \forall x:\ believes(a, \boldsymbol{\phi}),$
$type(x, b) \Rightarrow\ believes(x, \boldsymbol{\phi})$

> **"*All Republicans believe Trump and Trump believes he won the election.*" ⟹ "*Republicans believe they won the election.*"**

Introduction of the doxastic logic or the logic of context might lead to undecidability

What we want to model: Sentiments

Affective states toward something or someone

Sentiments are omnipresent in social interactions, especially between a company and its customers

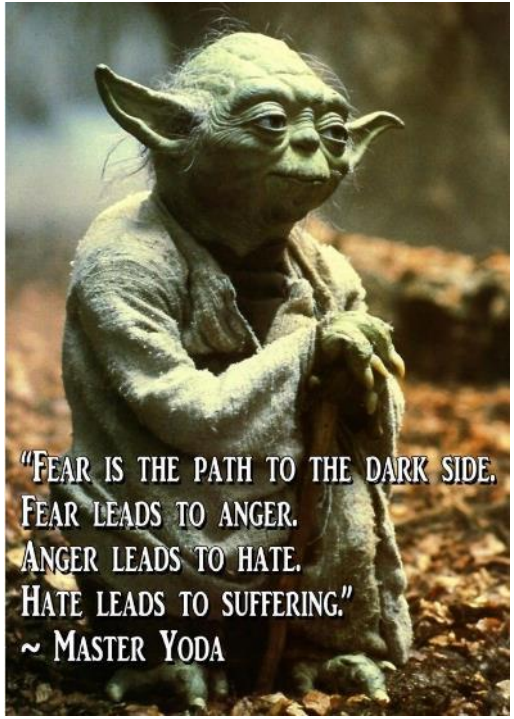
$\forall x: \neg like(x, a) \implies \neg buy(x, i)$

"John don't like Apple" ⟹ "He won't buy an iPhone"

$\forall x: like(x, h) \implies \psi(x)$

"John like to hunt" ⟹ "He won't vote for a political ecologist party"

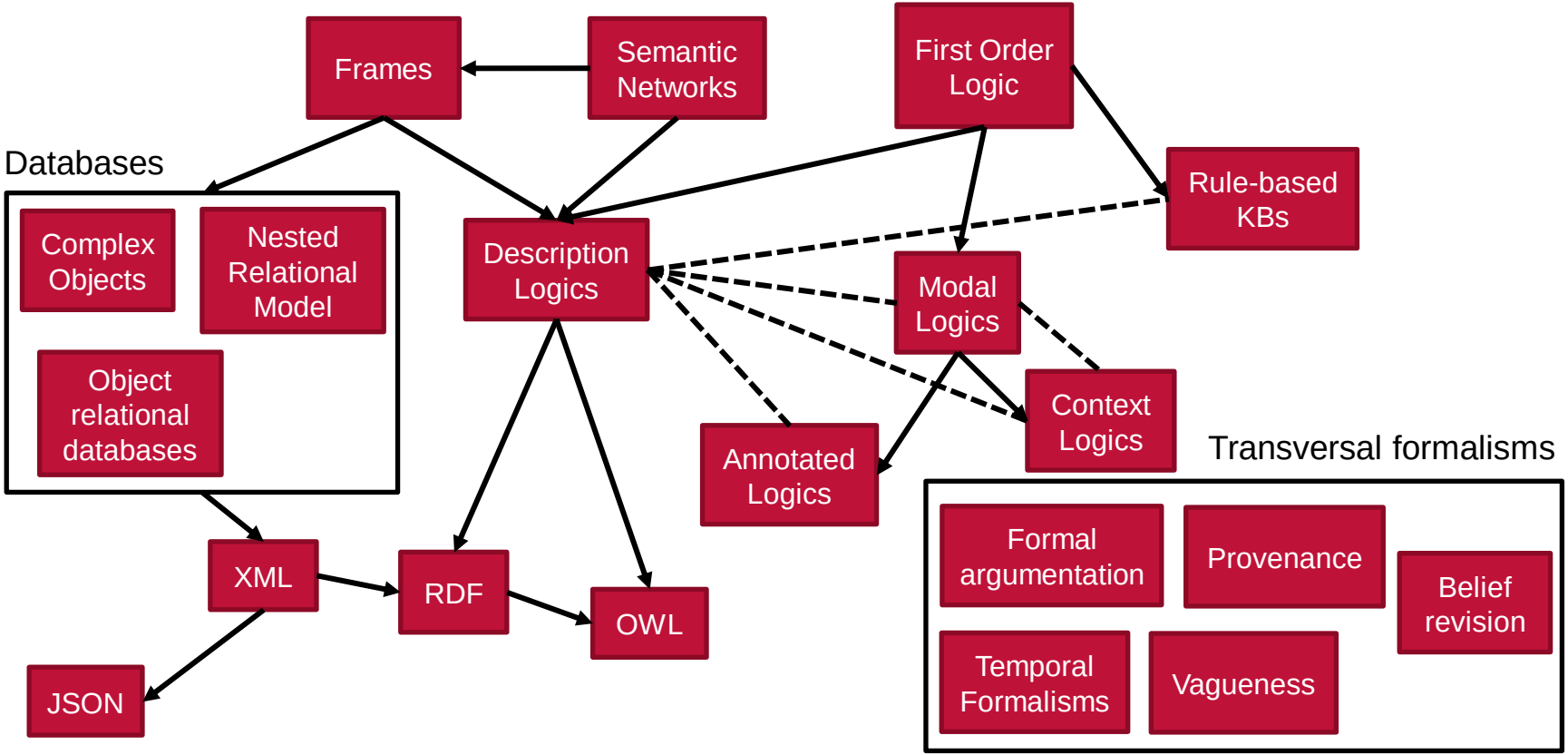
How to represent sentiments in knowledge bases?



What we want to model: Narratives

- **A series of logically and chronologically related events that describe something**
- **Narratives are everywhere: Wikipedia articles, newspapers, books,…**
- **“*Once upon a time…*”**
- **All of the preceding elements are necessary to properly represent and reason with narratives**

Existing Formalisms: A bestiary



Existing Formalisms: Challenges

- **What do we want?**

  To model and reason with high arity relationship, events, precedence, causation, negation, beliefs, sentiments and narratives

- **Some formalisms can represent what we need, but they are undecidable.**

- **We need a trade-off**