# PATTY

## Seminar Information Extraction
## Wednesday, 11th December 2013

Ezequiel Primo
Winter Semester 2013/14
Albert-Ludwigs-Universität Freiburg

UNI FREIBURG

# Outline

- Systems using IE with relevant tasks
- A taxonomy of relational patterns with semantic types
- Systematically harvest textual patterns from text corpora

# Informal Definition

- Large resource of relational patterns
- Large collection of relations
- Arranged by synonyms and into subsumptions
- Collection of relations learned automatically from text

# Glossary

- Entity: perceived or known or inferred to have its own distinct existence (living or nonliving)
- Subsume: classify, include, or incorporate in a more comprehensive category or under a general principle
- Thesaurus: list of words and terms utilized in order to represent certain concepts
- Synsets: stores semantically relationships between sets of synonyms

# Problem Statement

- Relationships between entities are expressed merely in latent form on the Web
- A priori unknown and need to be discovered in an unsupervised manner
- Comprehensively gathering and systematically organizing patterns for an open set of relations

# Motivation

UNI FREIBURG

- Automatically mining new relations from the Web
- Relationships between entities are expressed in highly diverse and noisy forms in natural-language text
- Approaches can detect and disambiguate named entities in text or tables

# Overview

- Patterns are organized into synonyms and subsumptions
- Patterns are semantically typed and organized into a taxonomy
- Extract binary relations between entities based on patterns in textual or semistructured contents
- Organizes a huge number of relational patterns into sets of synonymous patterns, and into a subsumption hierarchy

# Interpretation

Make use of a generalized notion of ontologically typed patterns

What is a pattern?

**(X) relationship with (Y)**

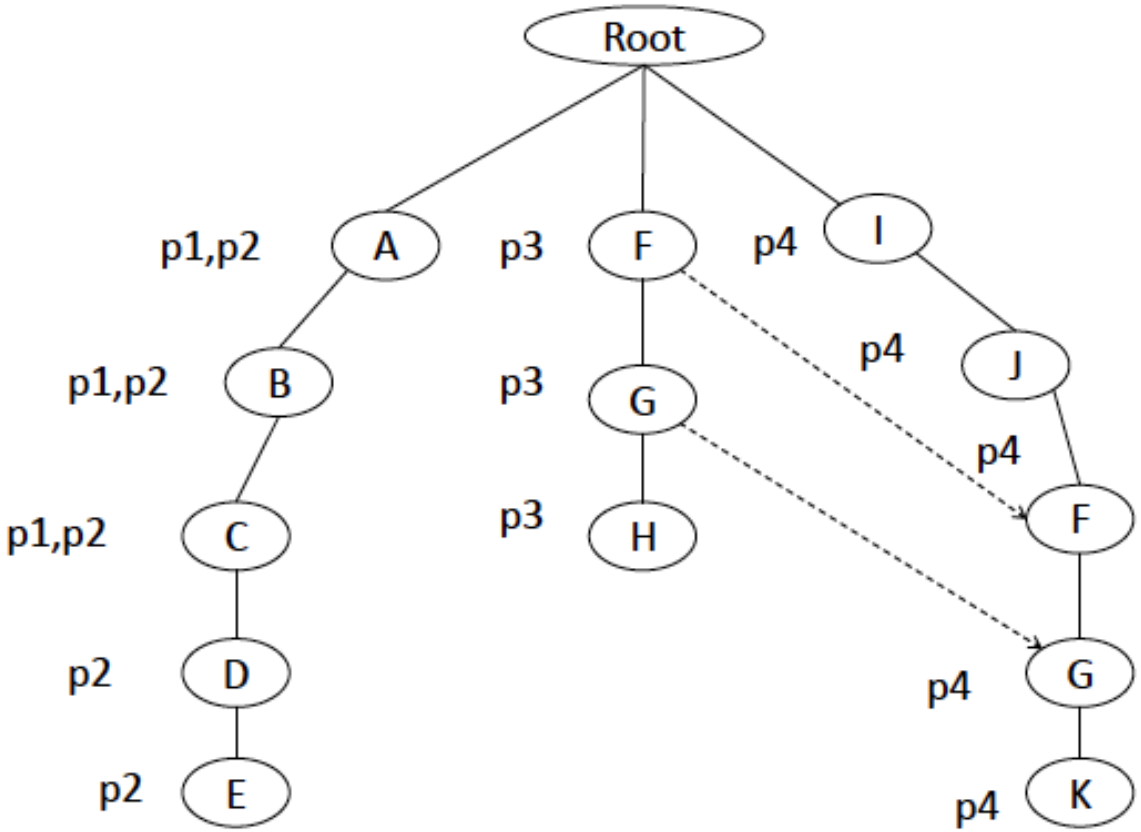
Samples:

**(Politician) was governor of (State)**

**(Actor × Movie)**

# Patterns Taxonomy

The patterns are semantically typed and organized into a subsumption taxonomy



# Relations

UNI FREIBURG

- Instances are always leaf (terminal) nodes in their hierarchies

- All noun hierarchies ultimately go up the root node (entity)

**Hyperonym:** The super-subordinate relation

**Meronymy:** The part-whole relation holds between synsets. E.g. {chair} and {back, backrest}

# Examples

UNI FREIBURG

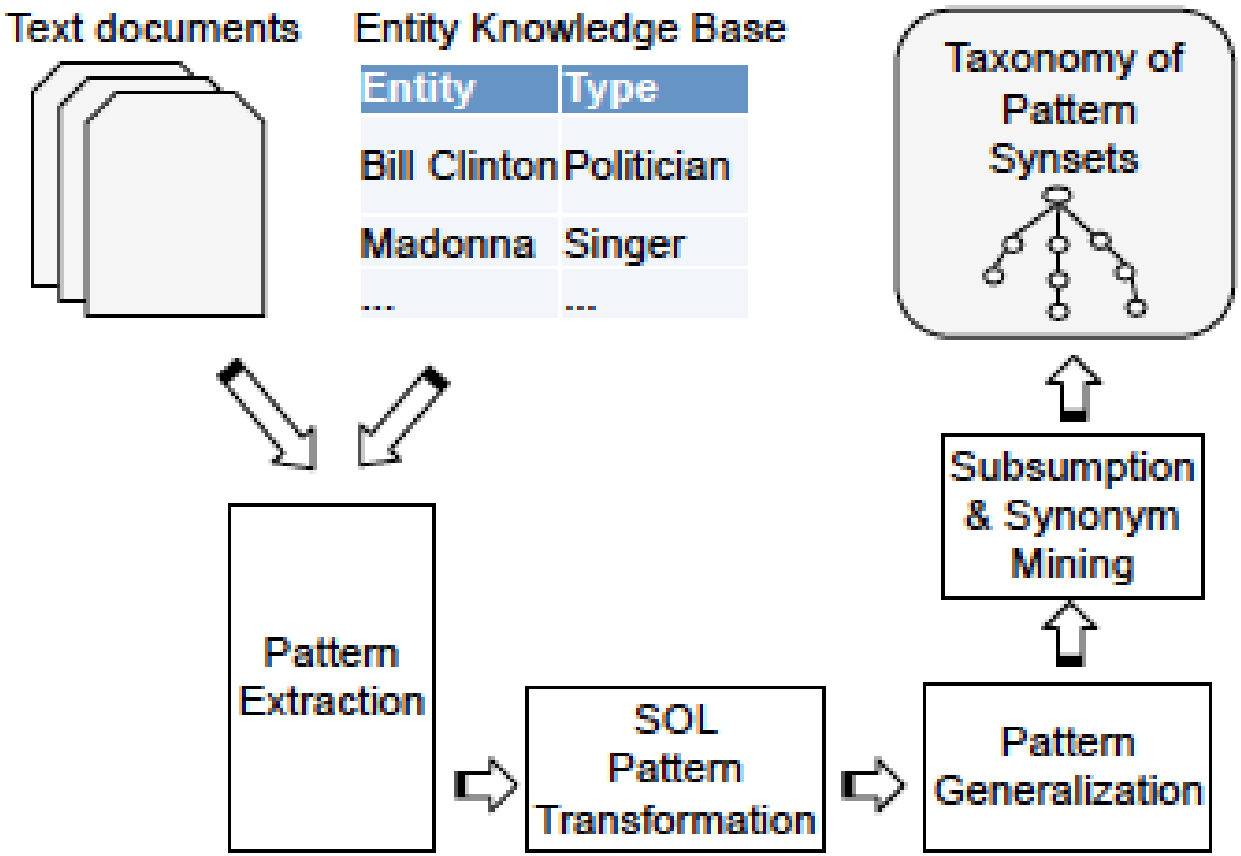
- **hasAcademicAdvisor** relation, simply by typing "worked under" into the search box
- **hasWonPrize** relation, was honored with, receive
  entity pairs such as (Bill Clinton, Arkansas), (Janet Napolitano, Arizona)
- Generalize
  actor _ award and musician _ award → person _ award
- **(Amy,Rehab)** → "Amy sings Rehab"

# Implementation

UNI FREIBURG

- Implemented in Java
- Hadoop as the platform for large-scale text and data analysis
- MongoDB for storing all resulting data
- Stanford NLP tool suite for linguistic processing
- AJAX for asynchronous communication with the server

# System Overview



# Architecture 1 - 2

UNI FREIBURG

1) Pattern extraction

- Output of pattern extraction are patterns extracted from paths of grammatical dependency graphs
- This information is used later for transforming basic patterns into SOL patterns

2) Pattern typing

- Generate typed patterns by attaching a type signature to each pattern
- Obtain the semantic types of the entities from our knowledge base

# Architecture SOL

- Transform plain patterns into syntactic-ontological lexical patterns
- Decompose textual patterns into n-grams (n consecutive words)
- Generate type signatures for these n-gram patterns
- Only frequent n-grams in corpus are retained in the SOL patterns, the rest replaced by wild-cards

# Architecture 3 - 4

UNI FREIBURG

3) Synset generation

- Generalize the patterns, both syntactically and semantically
- We group synonymous typed patterns into synonym sets, pattern synsets

4) Subsumption and synonym mining

- Relations between pattern synsets
- Arrange patterns into groups of synonyms and in a hierarchy based on hyperonym relations between patterns

# Roadmap

UNI FREIBURG

- Part of The YAGO-NAGA project
- Huge semantic knowledge base with almost 10 million entities (e.g. persons, organizations, cities)
- Derived from Wikipedia, WordNet and GeoNames
- Comprehensive repository of human knowledge
- Pre-defined knowledge base as input

# Usage

- Relation taxonomies from three different corpora
- Taxonomy derived from Wikipedia contains over 350,000 pattern synsets (June 2011 Version)
- The New York Times version contains ~ 80, 000 pattern synsets

# Assumptions

- The graph of subsumption relations might contain cycles, which have to be eliminated
- SOL methodology
- Antonyms, synonyms
- Synsets

# Applications

UNI FREIBURG

- Discover relationships between entities
- Assessing the truthfulness of search results or statements in social media
- Understanding in the challenging task of question answering
- Support automatic text analysis and IA applications

# Benefits

- Can boost IE and knowledge-base population tasks by its rich and clean repository of paraphrases for the relations
- Improve Open IE by associating type signatures with patterns
- Produce a combination of dictionary and thesaurus, which usage might be more intuitive

# Performance analysis

UNI FREIBURG

## Disadvantages

- The most expensive part is pattern extraction
- Limited to certain complex queries

## Advantages

- Has implicitly summarized the input text documents, users can exploit and query these summaries
- Different from existing systems, the user does not have to know the schema of the database

# Training

- We can identify subsumptions between relational patterns
  **“covered” is subsumed by “performed”**

- Discriminate patterns that can denote different relations based on their type
  **“covered” can refer to** musician _ song **or to** journalist _ event

- The user can restrict the domain of the relation
  Results with domains (music example)

# Examples Training

- To see what PATTY knows about Albert Einstein in his role as a scientist
- User can type Natalie Portman ?r ?x, Mila Kunis ?s ?x. This will find all entities **?x**
- PATTY finds the movie "Black Swan" for **?x** , and says that both actresses appeared in this movie
- Tom Cruise and Nicole Kidman are related, it is sufficient to type

# Demo

UNI FREIBURG

- User can enter Subject-Predicate-Object triples
- Users can browse the different meanings of patterns, as they occur with different types of entities
- Subsumption hierarchy of patterns, where more general patterns subsume more specific patterns

https://d5gate.ag5.mpi-sb.mpg.de/pattyweb/

# Evaluation and Results

UNI FREIBURG

- Subsumptions have a sampling-based accuracy of 83%
- 85% of the patterns are correct in the sense that they denote meaningful relations
- Wikipedia consists of about 350,000 typed-pattern synsets organized in a hierarchy with 8,162 subsumptions
- Music domain Out of 169 ground-truth relations, PATTY contains 126

# Comparisons

- ReVerb patterns for Open IE are fairly noisy and connect noun phrases rather than entities
- NELL is limited to a few hundred pre-specified relations
- None of the prior approaches knows the ontological types of patterns

# Summary and Trends

- PATTY's patterns contribute added value beyond today's knowledge bases
- Enriching Web data (both text and tables)
- Extract antonyms, where one relation is the opposite of another
- Future research inspire new applications in information extraction, QA, and text understanding

# References


- http://www.mpi-inf.mpg.de/yago-naga/patty/
- http://wordnet.princeton.edu/
- http://www.geonames.org/
- http://www.nytimes.com/ref/membercenter/nytarchive.html
- http://lemurproject.org/clueweb09/
- http://nlp.stanford.edu/
- http://www.freebase.com/
- http://yago-knowledge.org
- http://dbpedia.org
- http://reverb.cs.washington.edu
- http://en.wikipedia.org/wiki/English_language

# End

UNI FREIBURG

Thank you very much for your attention


Questions?