Large-scale information extraction from textual definitions through deep syntactic and semantic analysis

Andy Zhang

Abstract

- Present DEFIE, an approach to largescale Information Extraction (IE) based on a syntactic-semantic analysis of textual definitions.

  (textual definitions: short and concise descriptions of a given concept or entity)

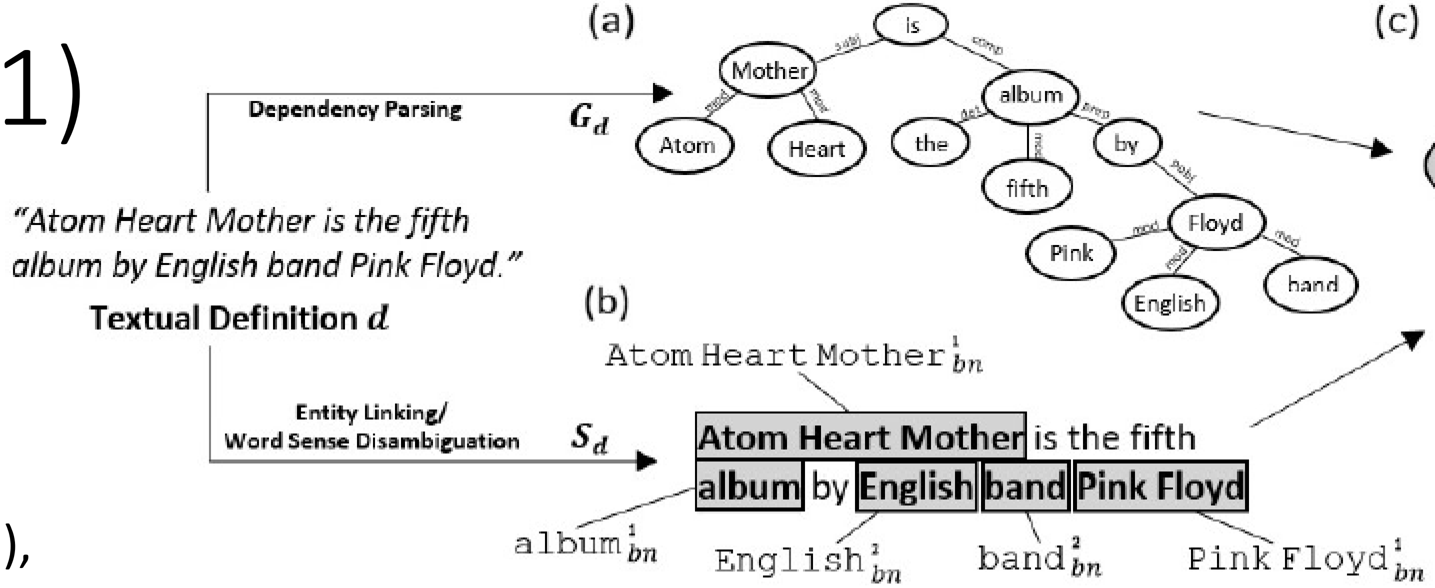
  - Leverage syntactic dependencies to reduce data sparsity
  - Disambiguate arguments & content words of the relation strings
  - Use the resulting info to organize the acquired relations hierarchically
- Output a knowledge base consisting of several million automatically acquired semantic relations

Shortcomings of previous works

- Constrained to small and often pre-specified sets of relations
- Rely mostly on dependencies at the level of surface text
- Relations strings are bound to surface text, lacking actual semantic content
- Require additional processing steps to be used in real applications

Relation extraction(1)

- Textual definition processing
  - Syntactic analysis -G_d
    - Parsing
    - using C&C (Clark and Curran, 2007), a log-linear parser based on Combinatory Categorial Grammar (CCG).
  - Semantic analysis -S_d
    - Based on Babelfy (Moro et al., 2014)
    - An approach to entity linking and word sense disambiguation

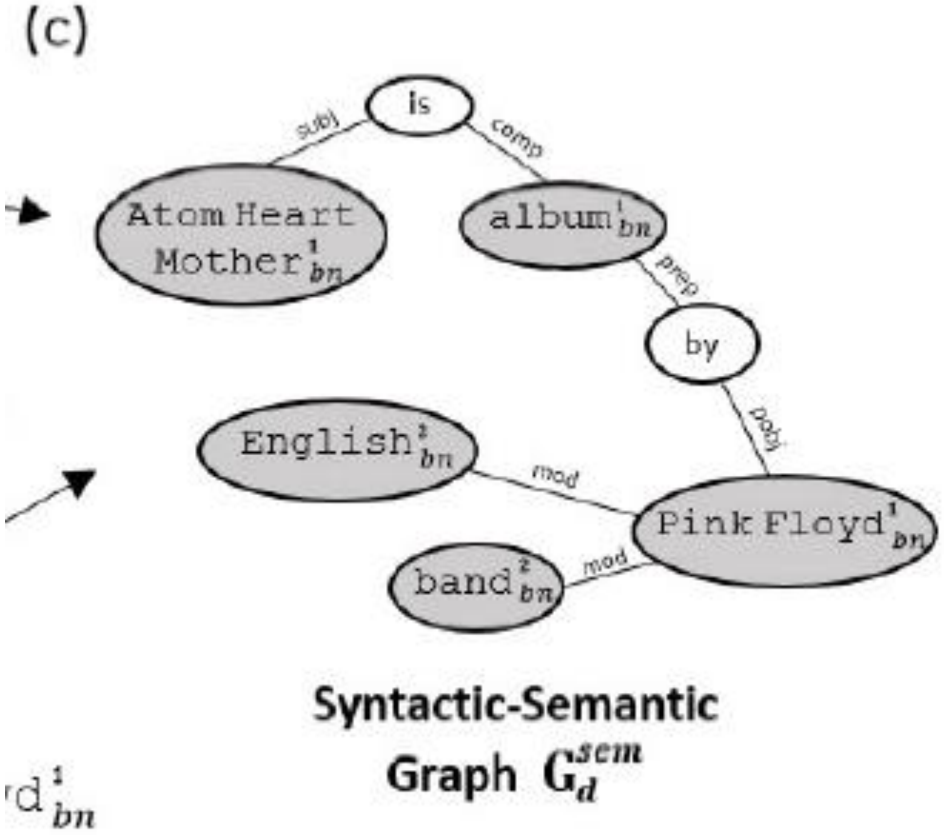


(a)
(c)
Dependency Parsing
G_d
is
Mother
album
Atom
Heart
the
by
fifth
Pink
Floyd
English
band
subj
comp
mod
det
prep
pobj
"Atom Heart Mother is the fifth album by English band Pink Floyd."
Textual Definition d
(b)
Entity Linking/ Word Sense Disambiguation
S_d
Atom Heart Mother$^1_{bn}$
Atom Heart Mother is the fifth album by English band Pink Floyd
album$^1_{bn}$
English$^1_{bn}$
band$^2_{bn}$
Pink Floyd$^1_{bn}$

Semantics draws on BabelNet(Navigli and Ponzetto, 2012)

Relation extraction(2)

- Syntactic-semantic graph construction
  - Merge vertices referring to same concept or entity
  - Incorporate semantic info from sense mapping S_d to vertices in dependency graph G_d
  - Discard non-disambiguated adjuncts and modifiers



Syntactic-Semantic Graph G_d^{sem}

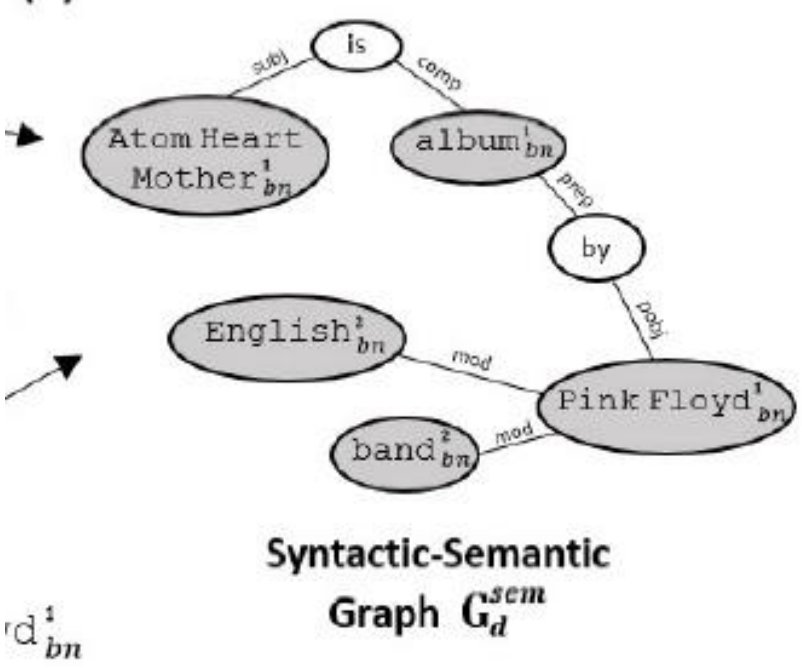
Relation extraction(3)

- Relation pattern identification
  - extract the relation pattern r between two entities and/or concepts as the shortest path between the two corresponding vertices in G_d^{sem}
  - Floyd-Warshall algorithm(Floyd, 1962)
  - One constraint: at least one verb

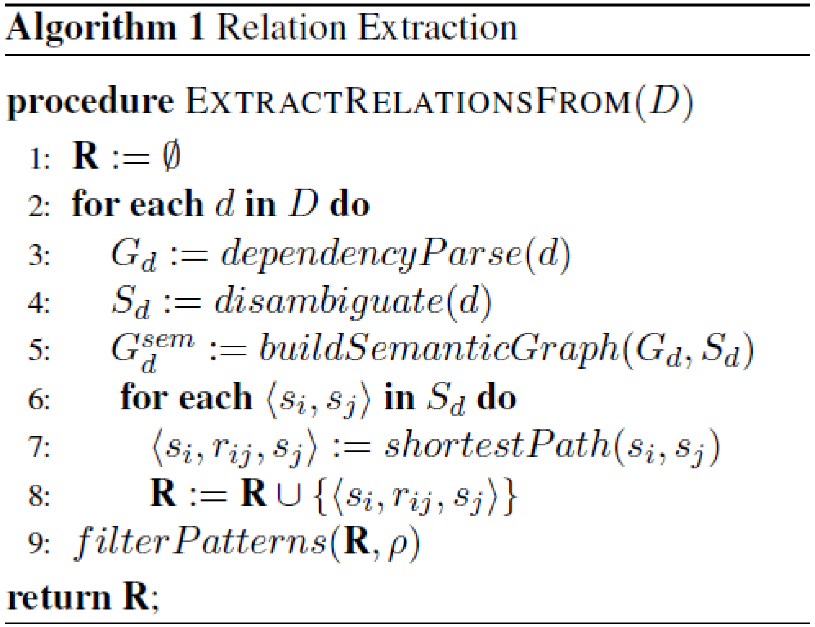
$$X \rightarrow \mathtt{is} \rightarrow \mathtt{album}^1_{bn} \rightarrow \mathrm{by} \rightarrow Y$$

$$X \rightarrow \mathtt{is} \rightarrow Y$$

(c)



Syntactic-Semantic Graph G_d^{sem}



Algorithm 1 Relation Extraction

```
procedure EXTRACTRELATIONSFROM($D$)
    $\mathbf{R} := \emptyset$
    for each $d$ in $D$ do
        $G_d := dependencyParse(d)$
        $S_d := disambiguate(d)$
        $G_d^{sem} := buildSemanticGraph(G_d, S_d)$
        for each $\langle s_i, s_j \rangle$ in $S_d$ do
            $\langle s_i, r_{ij}, s_j \rangle := shortestPath(s_i, s_j)$
            $\mathbf{R} := \mathbf{R} \cup \{\langle s_i, r_{ij}, s_j \rangle\}$
    $filterPatterns(\mathbf{R}, \rho)$
return R;
```

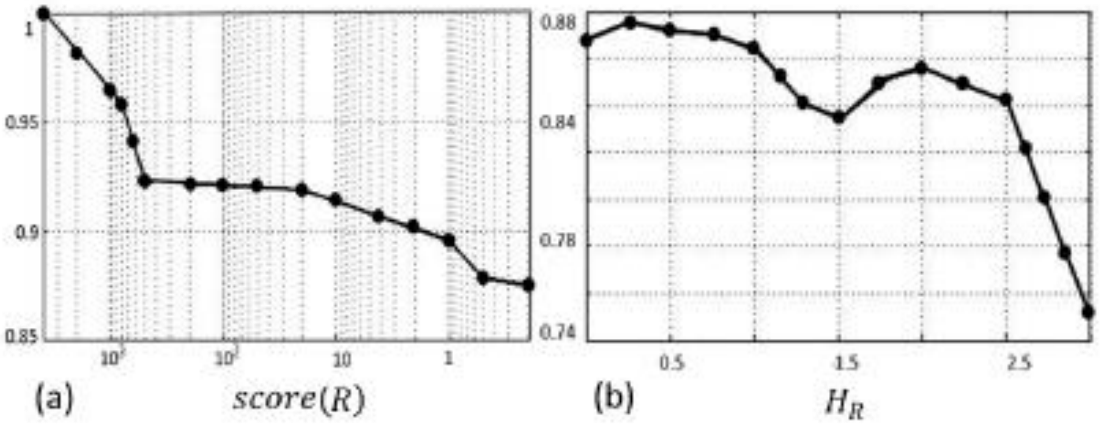
Relation type signatures and scoring

- Computing semantic type signatures for each relation
  - Collect hypernyms(BabelNet) of all the arguments, the one covers the biggest subset of arguments is selected to be the semantic class of the relation
- Scoring

$$\mathbf{R} := \mathbf{R} \cup \{\langle s_i, r_{ij}, s_j \rangle\}$$

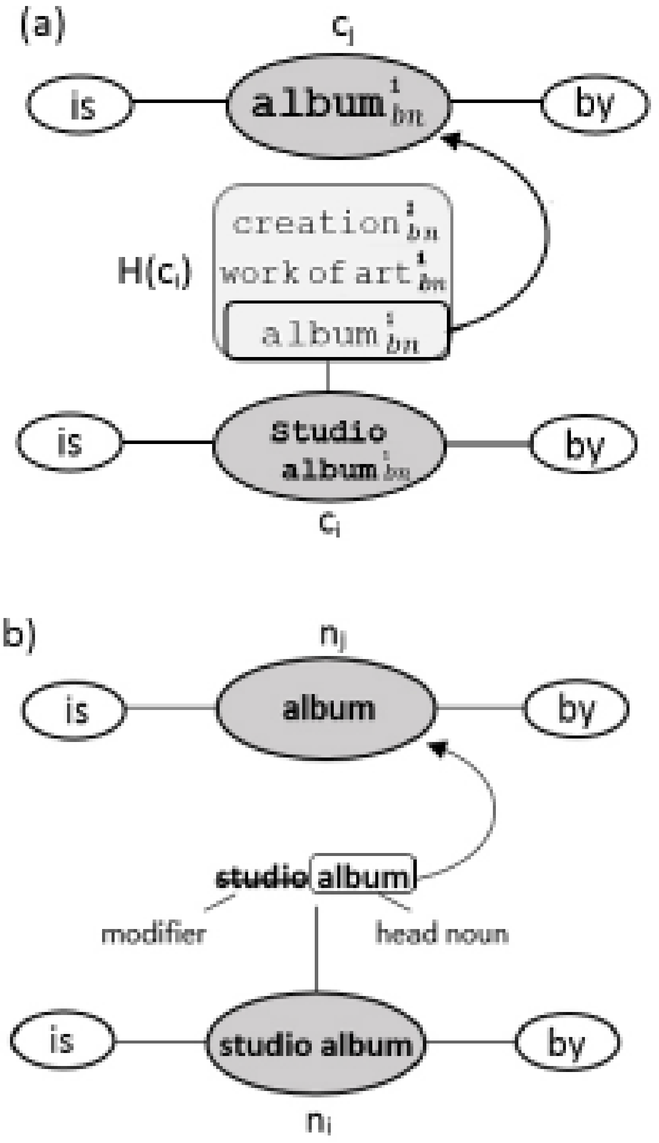
$$H_R = -\sum_{i=1}^{n} p(h_i)\, log_2\, p(h_i)$$

$$score(R) = \frac{|S_R|}{(H_R + 1)\, length(r)}$$



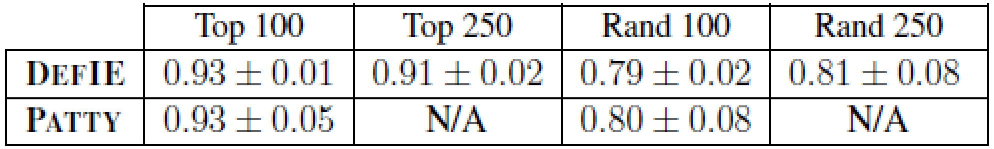
Relation taxonomization

- Consider only relations whose patterns are identical except for a single noun node
- Hypernym generalization
  - extract hypernym sets of concepts or entities
  - check whether one concept belongs to the set of the other
- Substring generalization



Experiment(1)

- All experiments conducted manually
- Assess the quality of relations
  - whether it represented a meaningful relation
  - whether the extracted argument pairs were consistent with this relation and the corresponding definitions

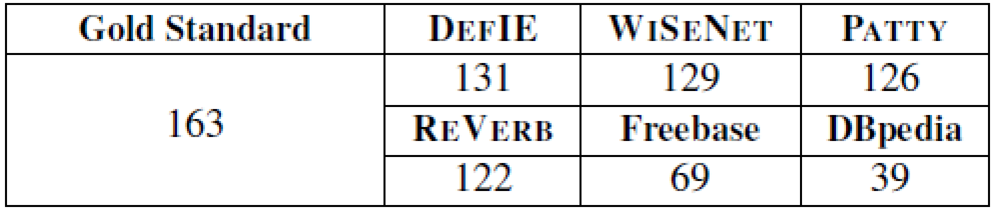


	Top 100	Top 250	Rand 100	Rand 250
DefIE	0.93 ± 0.01	0.91 ± 0.02	0.79 ± 0.02	0.81 ± 0.08
Patty	0.93 ± 0.05	N/A	0.80 ± 0.08	N/A

Table 3: Precision of relation patterns

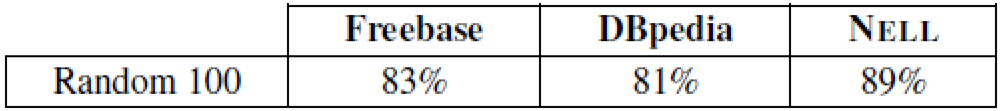
Experiment(2)

- Assess the coverage of relations
  - 163 manually annotated semantic relations from Wikipedia about musicians, seek for a relation carrying the same semantics



Gold Standard	DefIE	WiSeNet	Patty
163	131	129	126
	ReVerb	Freebase	DBpedia
	122	69	39

  - Look for similar relations in DEFIE

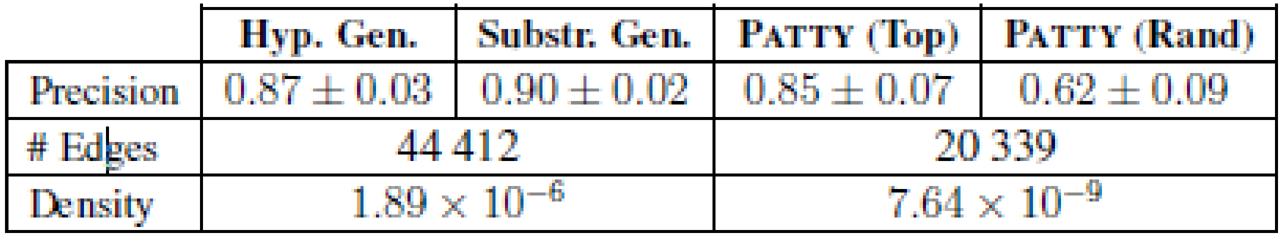


	Freebase	DBpedia	Nell
Random 100	83%	81%	89%

Table 6: Coverage of manually curated resources

Experiment(3)

- Quality of relation taxonomization
  - extracted a random sample of 200 hypernym edges for each generalization procedure
  - Manually judge whether they are correct or not



	Hyp. Gen.	Substr. Gen.	PATTY (Top)	PATTY (Rand)
Precision	0.87 ± 0.03	0.90 ± 0.02	0.85 ± 0.07	0.62 ± 0.09
# Edges	44 412		20 339	
Density	1.89×10^{-6}		7.64×10^{-9}	

Table 8: Precision and coverage of the relation taxonomy