

YAGO - Yet Another Great Ontology

YAGO: A Large Ontology from Wikipedia and WordNet¹

Presentation by: Besnik Fetahu

UdS

February 22, 2012

¹Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum

- 1 Introduction & Background
 - Ontology-Review
 - Usage of Ontology
 - Related Work
- 2 The YAGO Model
 - Aims of YAGO
 - Representation Models
 - Semantics
- 3 Putting all at work
 - Information Extraction Approaches
 - Quality Control
- 4 Evaluation & Examples
- 5 Conclusions

- 1 Introduction & Background
 - Ontology-Review
 - Usage of Ontology
 - Related Work
- 2 The YAGO Model
 - Aims of YAGO
 - Representation Models
 - Semantics
- 3 Putting all at work
 - Information Extraction Approaches
 - Quality Control
- 4 Evaluation & Examples
- 5 Conclusions

Ontology - Review

Knowledge represented as a set of concepts within a domain, and a relationship between those concepts.²

In general it has the following components:

- Entities
- Relations
- Domains
- Rules
- Axioms, etc.

In computer science and information science, an ontology formally represents knowledge as a set of concepts within a domain, and the relationships between those concepts. It can be used to reason about the entities within that domain and may be used to describe the domain. In theory, an ontology is a "formal, explicit specification of a shared conceptualisation".[1] An ontology renders shared vocabulary and taxonomy which models a domain with the definition of objects and/or concepts and their properties and relations.

Ontology (from onto-, from the Greek ὄν, ὄντος "being; that which is", present participle of the verb εἶμι "be", and -λογία, -logia: science, study, theory) is the philosophical study of the nature of being, existence, or reality as such, as well as the basic categories of being and their relations. Traditionally listed as a part of the major branch of philosophy known as metaphysics, ontology deals with questions concerning what entities exist or can be said to exist, and how such entities can be grouped, related within a hierarchy, and subdivided according to similarities and differences.

²[http://en.wikipedia.org/wiki/Ontology_\(information_science\)](http://en.wikipedia.org/wiki/Ontology_(information_science))

Ontology - Review

Knowledge represented as a set of concepts within a domain, and a relationship between those concepts.²

In general it has the following components:

- Entities
- Relations
- Domains
- Rules
- Axioms, etc.

In **computer science** and information science, an **ontology** formally represents knowledge as a set of **concepts** within a **domain**, and the **relationships** between those concepts. It can be used to reason about the entities within that domain and may be used to describe the domain. In theory, an ontology is a "formal, explicit specification of a shared conceptualisation".^[1] An ontology renders shared vocabulary and taxonomy which models

Ontology (from onto-, from the Greek ὄν, ὄντος "being; that which is", present participle of the verb εἶμι "be", and -λογία, -logia: science, study, theory) is the philosophical study of the nature of **being, existence**, or **reality** as such, as well as the basic categories of being and their relations. Traditionally listed as a part of the major branch of **philosophy** known as **metaphysics**, ontology deals with questions concerning what entities exist or can be said to exist, and how such entities can be grouped, related within a hierarchy, and subdivided according to similarities and differences.

²[http://en.wikipedia.org/wiki/Ontology_\(information_science\)](http://en.wikipedia.org/wiki/Ontology_(information_science))

Usage of Ontology (Wikipedia & WordNet)

Used in numerous fields of Semantic Web, and other fields like:

- Machine Translation
 - WordNet's synsets to resolve pattern disambiguity, for sentence translation (Chatterjee et al., 2005).

Usage of Ontology (Wikipedia & WordNet)

Used in numerous fields of Semantic Web, and other fields like:

- Machine Translation
 - WordNet's synsets to resolve pattern disambiguity, for sentence translation (Chatterjee et al., 2005).
- Word Sense Disambiguation
 - Wikipedia's categories, hyperlinks, and disambiguous articles, used to create a dataset of named entity (Bunescu R., and Pasca M., 2006).

Usage of Ontology (Wikipedia & WordNet)

Used in numerous fields of Semantic Web, and other fields like:

- Machine Translation
 - WordNet's synsets to resolve pattern disambiguity, for sentence translation (Chatterjee et al., 2005).
- Word Sense Disambiguation
 - Wikipedia's categories, hyperlinks, and disambiguous articles, used to create a dataset of named entity (Bunescu R., and Pasca M., 2006).
- Query Expansion
 - WordNet's synsets used to create different synonyms, hyponyms of a query term (Liu et al., 2004).

Usage of Ontology (Wikipedia & WordNet)

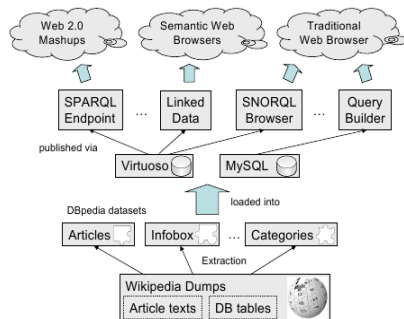
Used in numerous fields of Semantic Web, and other fields like:

- Machine Translation
 - WordNet's synsets to resolve pattern disambiguity, for sentence translation (Chatterjee et al., 2005).
- Word Sense Disambiguation
 - Wikipedia's categories, hyperlinks, and disambiguous articles, used to create a dataset of named entity (Bunescu R., and Pasca M., 2006).
- Query Expansion
 - WordNet's synsets used to create different synonyms, hyponyms of a query term (Liu et al., 2004).
- Document Classification.
 - WordNet's synsets used as concepts to create a link between word-concept to concept-document, in order to create a conditional probability distribution used later to classify documents (Ifrim G., and Weikum G., 2006).

List of current Ontologies

Ontologies built on data from Wikipedia, WordNet, or other available information on the web:

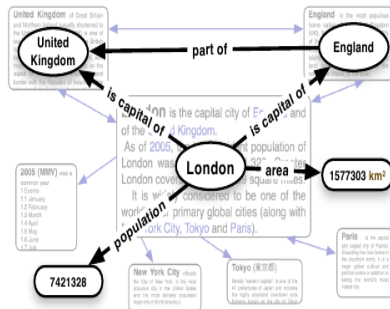
- KnowItAll
- TextRunner
- DBpedia
- SemanticWikipedia
- KYLIN, etc.



List of current Ontologies

Ontologies built on data from Wikipedia, WordNet, or other available information on the web:

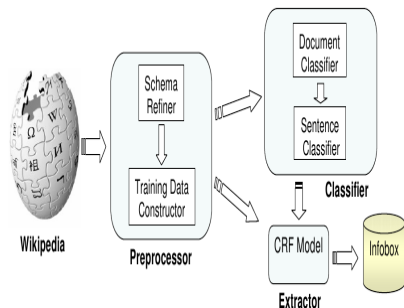
- KnowItAll
- TextRunner
- DBPedia
- SemanticWikipedia
- KYLIN, etc.



List of current Ontologies

Ontologies built on data from Wikipedia, WordNet, or other available information on the web:

- KnowItAll
- TextRunner
- DBPedia
- SemanticWikipedia
- KYLIN, etc.



Disadvantages of current Ontologies

Disadvantages and differences of previous ontologies and approaches with respect to YAGO:

- Non-Canonical Facts (TextRunner):
 - Different identifiers for same entity,
 - No clear relations,
 - No logic-based knowledge.
- Non-defined Relations, and Domains (DBPedia):
 - No ranges for relations
 - Word/Terms used from infoboxes appearing in different measuring metrics,
 - Consistency and accuracy thus unknown.
- Evaluation Results Missing (SemanticWikipedia):
 - No evaluation results, done for it (till now!).

- 1 Introduction & Background
 - Ontology-Review
 - Usage of Ontology
 - Related Work
- 2 The YAGO Model
 - Aims of YAGO
 - Representation Models
 - Semantics
- 3 Putting all at work
 - Information Extraction Approaches
 - Quality Control
- 4 Evaluation & Examples
- 5 Conclusions

Aims of YAGO

YAGO emerged as a need of creating a bigger ontology using current existing ontologies, its main aims were:

- Unification of Wikipedia and WordNet.



Aims of YAGO

YAGO emerged as a need of creating a bigger ontology using current existing ontologies, its main aims were:

- Unification of Wikipedia and WordNet.
- Make use of rich structures and information, such as: **Infoboxes, Category Pages**, etc.



Aims of YAGO

YAGO emerged as a need of creating a bigger ontology using current existing ontologies, its main aims were:

- Unification of Wikipedia and WordNet.
- Make use of rich structures and information, such as: **Infoboxes, Category Pages**, etc.

Subcategories

This category has the following 31 subcategories, out of 31 total.

- ▶ [Arts and culture templates](#) (24 C, 72 P)
- ▶ [Geography and place templates](#) (27 C, 88 P)
- ▶ [Health and fitness templates](#) (7 C, 7 P)
- ▶ [History and events templates](#) (24 C, 72 P)
- ▶ [Mathematics and abstraction templates](#) (7 C)
- ▶ [People and person templates](#) (21 C, 43 P)
- ▶ [Philosophy and thinking templates](#) (13 C, 16 P)
- ▶ [Religion and belief templates](#) (21 C, 95 P)
- ▶ [Science and nature templates](#) (20 C, 39 P)
- ▶ [Society and social science templates](#) (31 C, 52 P)
- ▶ [Sports templates](#) (16 C, 40 P)

Aims of YAGO

YAGO emerged as a need of creating a bigger ontology using current existing ontologies, its main aims were:

- Unification of Wikipedia and WordNet.
- Make use of rich structures and information, such as: **Infoboxes, Category Pages**, etc.
- Ensure plausibility of facts via type checking.

Subcategories

This category has the following 31 subcategories, out of 31 total.

- ▶ [Arts and culture templates](#) (24 C, 72 P)
- ▶ [Geography and place templates](#) (27 C, 88 P)
- ▶ [Health and fitness templates](#) (7 C, 7 P)
- ▶ [History and events templates](#) (24 C, 72 P)
- ▶ [Mathematics and abstraction templates](#) (7 C)
- ▶ [People and person templates](#) (21 C, 43 P)
- ▶ [Philosophy and thinking templates](#) (13 C, 16 P)
- ▶ [Religion and belief templates](#) (21 C, 95 P)
- ▶ [Science and nature templates](#) (20 C, 39 P)
- ▶ [Society and social science templates](#) (31 C, 52 P)
- ▶ [Sports templates](#) (16 C, 40 P)

YAGO Model

Representation Model in YAGO:

- *Data Model*: extension to RDFS, includes **acyclic transitivity (atr)**.
- *Entities*: abstract ontological objects, with the following properties
 - Each entity is part of at least one class,
 - Classes arranged in taxonomic hierarchy,
 - Relations are entities (express transitivity of relations - **atr**),
 - Facts are the triple: entity, relation, entity,
 - Each fact has an identifier.

YAGO Entities (n-ary) relation

fact:1 #1 Elvis Presley BornInYear 1935

fact:2 #2 #1 FoundIn Wikipedia

Elvis Presley BornInYear 1935 FoundIn Wikipedia

Semantics & Reification Graphs

YAGO's description through *Reification Graphs*:

Reification Graphs

- finite set of common entities \mathcal{C} .
- finite set of fact identifiers \mathcal{I} .
- finite set of relation names \mathcal{R} .

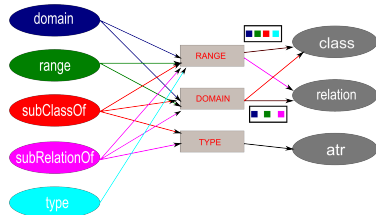
$$y : \mathcal{I} \rightarrow (\mathcal{I} \cup \mathcal{C} \cup \mathcal{R}) \times \mathcal{R} \times (\mathcal{I} \cup \mathcal{C} \cup \mathcal{R}) \quad (1)$$

- 1 Minimal set of common entities:
 $\mathcal{C} = \{\text{entity, class, relation, atr}\}$
- 2 Minimal set of relation names:
 $\mathcal{R} = \{\text{type, subclassOf, domain, range, subRelationOf}\}$

Semantics & Reification Graphs

Semantics and operations over the YAGO ontology, are defined as follows:

- 1 facts: $\mathcal{F} : \mathcal{I} \rightarrow (\mathcal{I} \cup \mathcal{C} \cup \mathcal{R}) \times \mathcal{R} \times (\mathcal{I} \cup \mathcal{C} \cup \mathcal{R})$
- 2 rewrite system: $F \cup \{f_1, \dots, f_n\} \rightarrow F \cup \{f_1, \dots, f_n\} \cup f, \forall F \subseteq \mathcal{F}$
- 3 axiomatic rules for the rewrite system:



$$\forall r, r_1, r_2 \in \mathcal{R}, x, y, c, c_1, c_2 \in \mathcal{I} \cup \mathcal{C} \cup \mathcal{R}, r_1 \neq \text{TYPE},$$

$$r_2 \neq \text{SubRelationOf}, r \neq \text{SubRelationOf}, r \neq \text{TYPE}$$

$$c \neq \text{atr}, c_2 \neq \text{atr}$$

- (1) $\{(r_1, \text{SubRelationOf}, r_2), (x, r_1, y)\} \leftrightarrow (x, r_2, y)$
- (2) $\{(r, \text{TYPE}, \text{atr}), (x, r, y), (y, r, z)\} \leftrightarrow (x, r, z)$
- (3) $\{(r, \text{DOMAIN}, c), (x, r, y)\} \leftrightarrow (x, \text{TYPE}, c)$
- (4) $\{(r, \text{RANGE}, c), (x, r, y)\} \leftrightarrow (y, \text{TYPE}, c)$
- (5) $\{(x, \text{TYPE}, c_1), (c_1, \text{SubClassOf}, c_2)\} \leftrightarrow (x, \text{TYPE}, c_2)$

Other Semantic Properties of YAGO

- 1 **Equivalency**: two ontologies $\mathbf{y}_1, \mathbf{y}_2$ are equivalent if the fact identifiers in \mathbf{y}_2 can be renamed by a bijective substitution.
- 2 **Consistency**: an ontology \mathbf{y} is called consistent **iff** there exists a model for it.
- 3 **Query language**: a pattern for a reification graph $G_{N,I,L}$ over a set of variables $V, V \cap (N \cup I \cup L) = \emptyset$ is a reification graph over set of nodes $N \cup V$, the set of identifiers $I \cup V$, and the set of labels $L \cup V$. A matching of a pattern P for a graph G is a substitution $\sigma : V \rightarrow N \cup I \cup L$, such that $\sigma(P) \subset G$, is called a match.

YAGO Model

- 1 *all facts of y are true in structure*
- 2 *if $\Psi(x, TYPE, string)$ for some x , then $\mathcal{D}(x) = x$*
- 3 *if $\Psi(r, TYPE, atr)$ for some r , then $\nexists x : \text{s.t. } \Psi(x, r, x)$*


- 1 Introduction & Background
 - Ontology-Review
 - Usage of Ontology
 - Related Work
- 2 The YAGO Model
 - Aims of YAGO
 - Representation Models
 - Semantics
- 3 Putting all at work**
 - Information Extraction Approaches
 - Quality Control
- 4 Evaluation & Examples
- 5 Conclusions

Data Sources

- ① **Wikipedia**: with **3,867,050** articles in english wikipedia (Feb. 2012), with over **2 billion** words³, whereas in YAGO's case was used the dataset (Nov. 2007), with **2,000,000** articles. Important attributes in the wikipedia articles are infoboxes, categories, etc.
- ② **WordNet**: a semantic lexicon for the english language, with its very useful data about synsets⁴. It contains, **82,115** synsets, and **117,798** nouns, etc.
- ③ **(+) More datasets**: As in the extension of YAGO2⁵ its used *geo-location* information for entities with permanent locations, i.e. cities, lakes, etc, taken from **Geonames**.

³ http://en.wikipedia.org/wiki/Wikipedia:Size_comparisons

⁴ Synset: a set of words that share one sense

⁵ YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages 

Information Extraction

In this part is established the connection between Wikipedia and WordNet, using the following IE techniques:

- **Wikipedia Heuristics:** each wikipedia page is considered a candidate to become an individual in YAGO.
- **Infobox Heuristics:** up to 170 valuable attributes found, for which are created manually YAGO relations.
- **Parsing:** convert literal values, and normalize them into ISO standards.

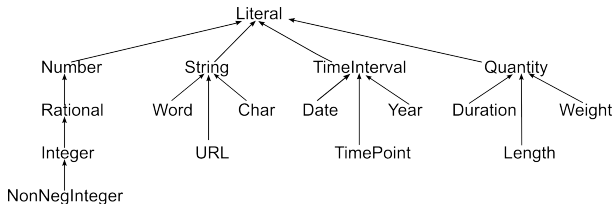


Figure: YAGO literal cases.

Types of Facts

To determine the class of each individual in YAGO, there are used different heuristics:

- 1 **Wikipedia Categories:** used to establish for each individual its class, the Wikipedia Category is used, i.e. *Albert Einstein*, is in the category of *Naturalized Citizens of US*.
- 2 **Conceptual Categories:** are candidates serving as a class for an individual.
- 3 **The Wikipedia Category Hierarchy:** they are organized in a directed acyclic graph, in which only the leafs of the graphs are used as possible candidates for an individual.
- 4 **WordNet Synsets:** common nouns in WordNet and Wikipedia are considered as classes.

Connecting Wikipedia and WordNet

The extracted classes are parsed into pre/post(-modifiers) and head compound, using *Noun Group Parser* (Suchanek F. et al., 2006), and the procedure is as follows:

Function `wiki2wordnet(c)`

Input: Wikipedia category name *c*

Output: WordNet synset

```
1  head =headCompound(c)
2  pre =preModifier(c)
3  post =postModifier(c)
4  head =stem(head)
5  If there is a WordNet synset s for pre + head
6    return s
7  If there are WordNet synsets s1, ..., sn for head
8    (ordered by their frequency for head)
9    return s1
10 fail
```

- \exists synset for pre-modifier, and head compound, then Wikipedia class becomes a subclass of the WordNet class.
- \nexists synset, map head compound to the WordNet synset with the highest frequency.

Other Heuristics on IE

- 1 **Synsets:** exploit the meaning of different classes from wikipedia, i.e. *metropolis*, and *urban center* both are equivalent to the synset *city*.
- 2 **Redirects:** using Wikipedia redirects to add valuable information to the relation MEANS, i.e. **Einstein, Albert MEANS Albert Einstein**.
- 3 **Person Names:** extract name components to add information to relations such as: GivenNameOf, FamilyNameOf.
- 4 **Relational Categories:** it adds valuable information from the Wikipedia Categories, i.e **Rivers in Germany**, can be extracted information such as LocatedIn, etc.
- 5 **Language Categories:** for some Wikipedia categories, exist the equivalent category name in different languages, thus is extracted information for relations such as: IsCalled, and InLanguage, i.e. **London IsCalled 'Londres' InLanguage French**.

Quality Control & Type Checking

The purpose of quality control was to deliver high quality ontology, and overcome the drawbacks of previous ontologies:

- **Canonicalization:** each fact and entity has a unique reference
 - **Redirect Resolution:** incorrect typed candidate facts, are resolved into correct ones, using Wikipedia redirects.
 - **Duplicate facts:** removes duplicate facts, and only the most precise are kept, i.e, birthday **1935-01-08** is favored instead of **1935**.
- **Type Checking:** checks the plausibility of generated facts:
 - **Reductive:** for a candidate fact that there couldn't be determined a class, is removed. While for facts that their class is not in the expected domain those are removed too.
 - **Inductive:** for entities with a birthdate in most cases those are persons, thus if they don't have a class, for those is assigned *person* as a class rather than eliminating them.

- 1 Introduction & Background
 - Ontology-Review
 - Usage of Ontology
 - Related Work
- 2 The YAGO Model
 - Aims of YAGO
 - Representation Models
 - Semantics
- 3 Putting all at work
 - Information Extraction Approaches
 - Quality Control
- 4 Evaluation & Examples
- 5 Conclusions

Relation Precision

	Heuristic	#Eval	Precision
1	hasExpenses	46	100.0% \pm 0.0%
2	hasInflation	25	100.0% \pm 0.0%
3	hasLaborForce	43	97.67441% \pm 0.0%
4	during	232	97.48950% \pm 1.838%
5	ConceptualCategory	59	96.94342% \pm 3.056%
6	participatedIn	59	96.94342% \pm 3.056%
7	plays	59	96.94342% \pm 3.056%
8	establishedInYear	57	96.84294% \pm 3.157%
9	createdOn	57	96.84294% \pm 3.157%
10	originatesFrom	57	96.84294% \pm 3.157%
	...		
72	WordNetLinker	56	95.11911% \pm 4.564%
	...		
74	InfoboxType	76	95.08927% \pm 4.186%
75	hasSuccessor	53	94.86150% \pm 4.804%
	...		
88	hasGDPPPPP	75	91.22189% \pm 5.897%
89	hasGini	62	91.00750% \pm 6.455%
90	discovered	84	90.98286 \pm 5.702%

Table: Precision of YAGO's heuristics

Largest Relations in YAGO

Relation	#Facts	Relation	#Facts
hasUTCOffset	12724	hasWonPrize	13645
livesIn	15 185	writtenInYear	16441
originatesFrom	16876	directed	18633
hasPredecessor	19154	actedIn	22249
hasDuration	23652	bornInLocation	24400
hasImdb	24659	hasArea	26781
hasProductLanguage	27840	produced	30519
hasPopulation	30731	isOfGenre	33898
hasSuccessor	46658	establishedOnDate	69529
hasWebsite	79779	created	83627
locatedIn	125738	diedOnDate	168037
subClassOf	211979	bornOnDate	350613
givenNameOf	464816	familyNameOf	466969
isLanguage	2389627	isCalled	2984362
type	3957223	means	4014819

Table: Largest relations in YAGO

YAGO Ontology usage Examples

Query	Result
Who was Einstein's doctoral advisor? Einstein <i>hasDoctoralAdvisor</i> ?x	?x = Alfred Kleiner
Who is named after a place in Africa? ?place <i>locatedIn</i> Africa ?name <i>means</i> ?place ?name <i>familyNameOf</i> ?who	who = Gabriel Sudan and 22 more

Table: Sample queries on YAGO

Id	Subject	Property	Object
1 #508944559	Albert Einstein	hasAcademicAdvisor	Alfred Kleiner
#2125828	Albert Einstein	means	Albert Einstein
2 #508954111	Albert Einstein	hasAcademicAdvisor	Alfred Kleiner
#2125828	Albert Einstein	means	Albert Einstein

Id	Subject	Property	Object
1 #524890131	Abba_Imo State	isLocatedIn	Africa
#327866390	Abba	means	Abba_Imo State
#378015012	Amar Abba	hasFamilyName	Abba
#96893054	Africa	means	Africa

- 1 Introduction & Background
 - Ontology-Review
 - Usage of Ontology
 - Related Work
- 2 The YAGO Model
 - Aims of YAGO
 - Representation Models
 - Semantics
- 3 Putting all at work
 - Information Extraction Approaches
 - Quality Control
- 4 Evaluation & Examples
- 5 Conclusions

Conclusions

Main features of YAGO, and its contributions:

- 1 High coverage and high quality ontology.
- 2 Integration of two largest ontologies Wikipedia, and WordNet.
- 3 Usage of structured information such as *Infoboxes*, *Wikipedia Categories*, *WordNet Synsets*.
- 4 Introduction of a new data model, *YAGO Model*.
- 5 Expression of acyclic transitive relations.
- 6 Type checking, ensuring that only plausible facts are contained.
- 7 Canonical facts.
- 8 Query Engine, etc.

YAGO's impact



Figure: YAGO and the relation to other Ontologies.

YAGO goes 2.0

- 1 New types of rules: **Factual rules, Implication rules, Replacement rules, Extraction rules**
- 2 Extracting information from different point of views:
 - 1 Temporal Dimension: Assign begin and/or end of time spans to all entries, facts, events, etc.
 - 2 Geo-Spatial Dimension: assign location in space to all entities having a permanent location.
 - 3 Textual Dimension: extract information from Wikipedia, for relation such as `hasWikipediaAnchorText`, `hasCitationTitle`, etc, and also multi-lingual is considered in this dimension.
- 3 80 million facts in YAGO2 with near-human quality.

Bibliography

- 1 YAGO: A Large Ontology from Wikipedia and WordNet - F. M. Suchanek, Gj. Kasneci, G. Weikum
- 2 YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages -
- 3 DBpedia: A Nucleus for a Web of Open Data - S. Auer, Ch. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives
- 4 Open Information Extraction from the Web - M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni
- 5 Using Encyclopedic Knowledge for Named Entity Disambiguation - R. Bunescu, M. Paşca
- 6 Resolving Pattern Ambiguity for English to Hindi Machine Translation Using WordNet - N. Chatterjee, Sh. Goyal, A. Naithani
- 7 Web-Scale Information Extraction in KnowItAll - O., M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. S. Weld, A. Yates
- 8 Transductive Learning for Text Classification Using Explicit Knowledge Models - G. Ifrim, and G. Weikum
- 9 An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases - S. Liu, C. Yu, W. Meng
- 10 Semantic Wikipedia - M. Völkel, M. Krötzsch, D. Vrandečić, H. Haller, R. Studer
- 11 Autonomously Semantifying Wikipedia - F. Wu, D. S. Weld

Thank you!
Questions?