New Trends

Pay as you go approach [Madhavan et al 2007]

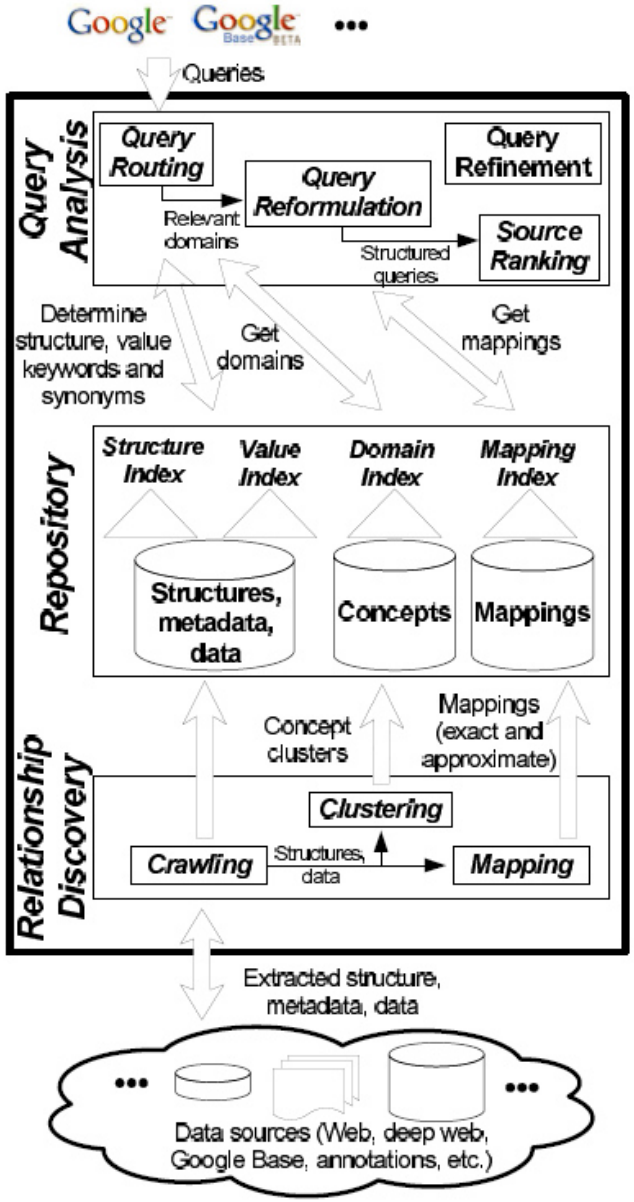
- There is no single mediated schema over which users pose queries.
  - There are sets of schemata that are clustered into topics
  - Semantic mappings are typically approximate
  - Queries are typically posed as keywords, respecting the main search paradigm on the Web, and are *routed* to the relevant sources
- The pay-as-you-go principle states that the system needs to be able to incrementally evolve its *understanding* of the data it encompasses as it runs
- Several automated methods to bootstrap the understanding of underlying data
  - the results of all these automated methods should be verified by humans, who can quickly correct the errors
  - At web scale, it is impossible for humans to inspect all of these results. It is crucial that we leverage as much as possible feedback we can get from users
- To support all of the above, a PAYGO-based system needs to model uncertainty at all levels:
  - queries,
  - mappings,
  - underlying data.

Pay as you go approach (2)

- The proposal is customized for structured data on the web
- The authors identify three scenarios:
  - The deep web
  - Google base
  - Annotation schemas
- The integration at the web-scale is different from the usual scenario: data on the web is about everything → there is no domain knowledge that may be used, heterogeneity is very high

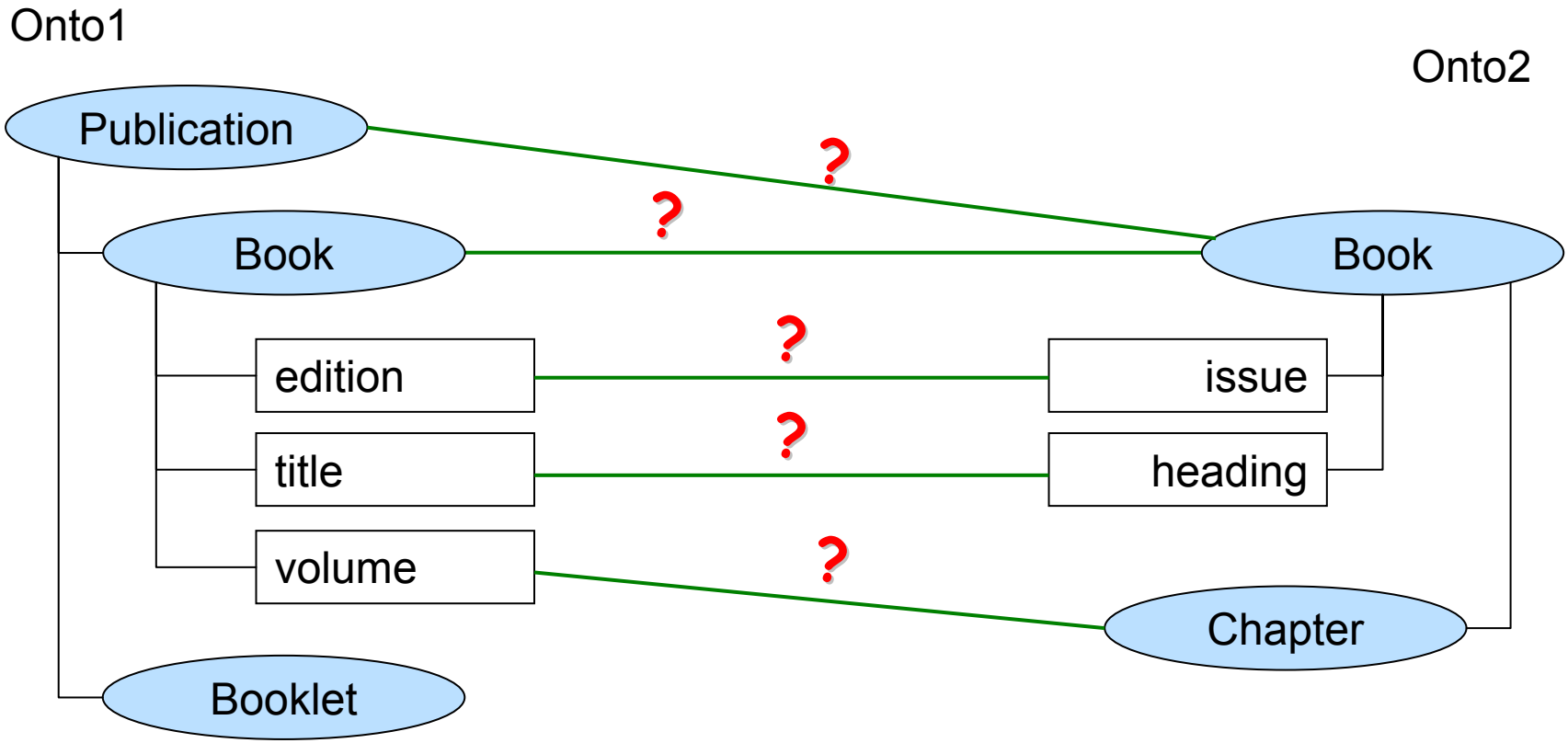
Traditional Data Integration	PAYGO Data Integration
Mediated Schema	Schema clusters
Schema mappings	Approximate mappings
Structured queries	Keyword queries with query routing
Query answering	Heterogeneous result ranking

Pay as you go architecture



ia &

Schema matching: a probabilistic approach



L. Po, S. Sorrentino, S. Bergamaschi, D. Beneventano –
Lexical Knowledge Extraction: an Effective Approach to Schema and Ontology Matching, Eckm 2009

Lexical knowledge inside sources

- Elements of schemata/ontologies are labelled by natural language expressions. Natural language labels provide a rich connection between formal objects (e.g. classes and properties) and their intended meanings

- It is necessary to address the problem of how the data are "labelled", i.e. understanding the meaning behind the names denoting schemata/ontology elements

What role Lexical Knowledge plays in data integration and ontology engineering?

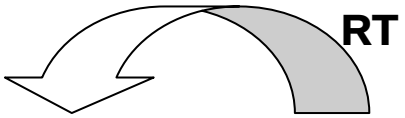
- Matching techniques aim at finding correspondences between semantically related entities of different schemata/ontologies
- We propose a matching technique based on **Automatic Lexical Annotation**:
  - Automatic Lexical Annotation of schemata/ontologies is performed
  - probabilistic lexical relationships are discovered

Lexical Annotation

- *Annotation* is a piece of information added to a book, document, online record, video, or other data
- *Lexical Annotation* is an annotation performed w.r.t. a *Semantic Resource* such as a thesaurus (for example Roget) or a Semantic Lexicon (like WordNet)
- Each annotation has the property to own a lexical description. *Lexical Annotation* differs from the *Ontology-based Annotation* where annotation is w.r.t. an ontology (top ontology or domain ontology)

Lexical Annotation – an example

- Lexical Annotation
  - assigns meanings to class and attribute names w.r.t. a semantic resource (WordNet)
  - derives relationships among terms of the sources



	Word form		
Meaning (synset)	*Book*	*Volume*	*Publication*
a written work or composition that has been published (printed on pages bound together)	√		
physical objects consisting of a number of pages bound together; "he used a large book as a doorstop"	√	√	
a copy of a printed work offered for distribution			√

Is a kind of

lexical relationships extracted

Book	SYN	Volume
Book	RT	Publication

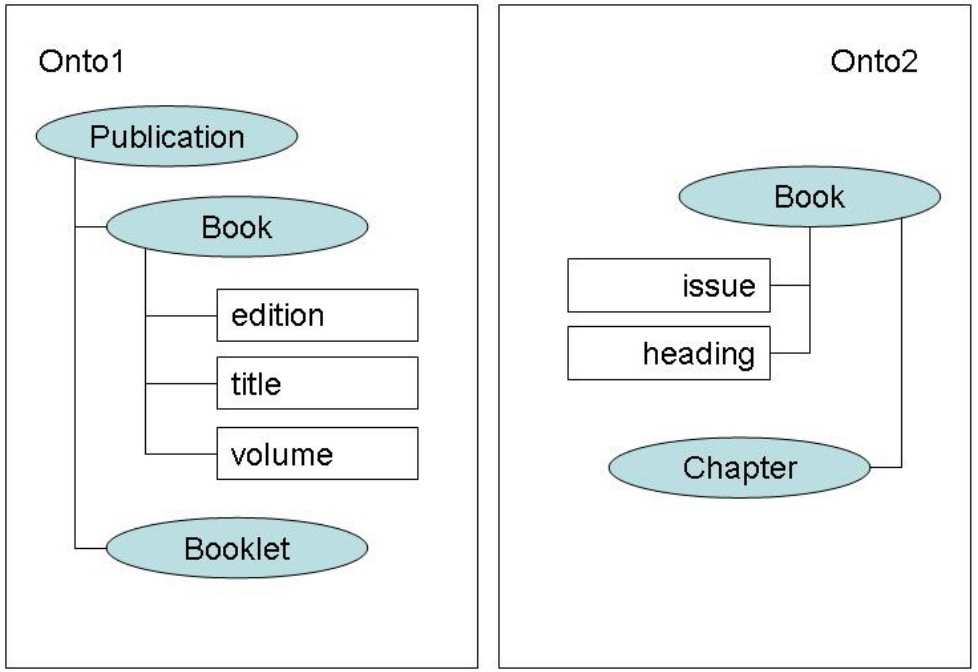
PWSD – Probabilistic Word Sense Disambiguation Algorithm [Bergamaschi et al. ECKM'09]

- Automatic lexical annotation is performed by the use of **Word Sense Disambiguation (WSD)** techniques
- To maximize annotation accuracy we need to employ a variety of WSD algorithms

- **PWSD** is a probabilistic method to combine the results of a set of WSD algorithms through the use of the Dempster's rule of combination (Shafer, 1976)
- PWSD automatically annotates terms of data sources and associates each annotation with a probability value that indicates the reliability level of the annotation
- PWSD has been implemented in the ALA (Automatic Lexical Annotation) tool [Bergamaschi et al. ER' 09]

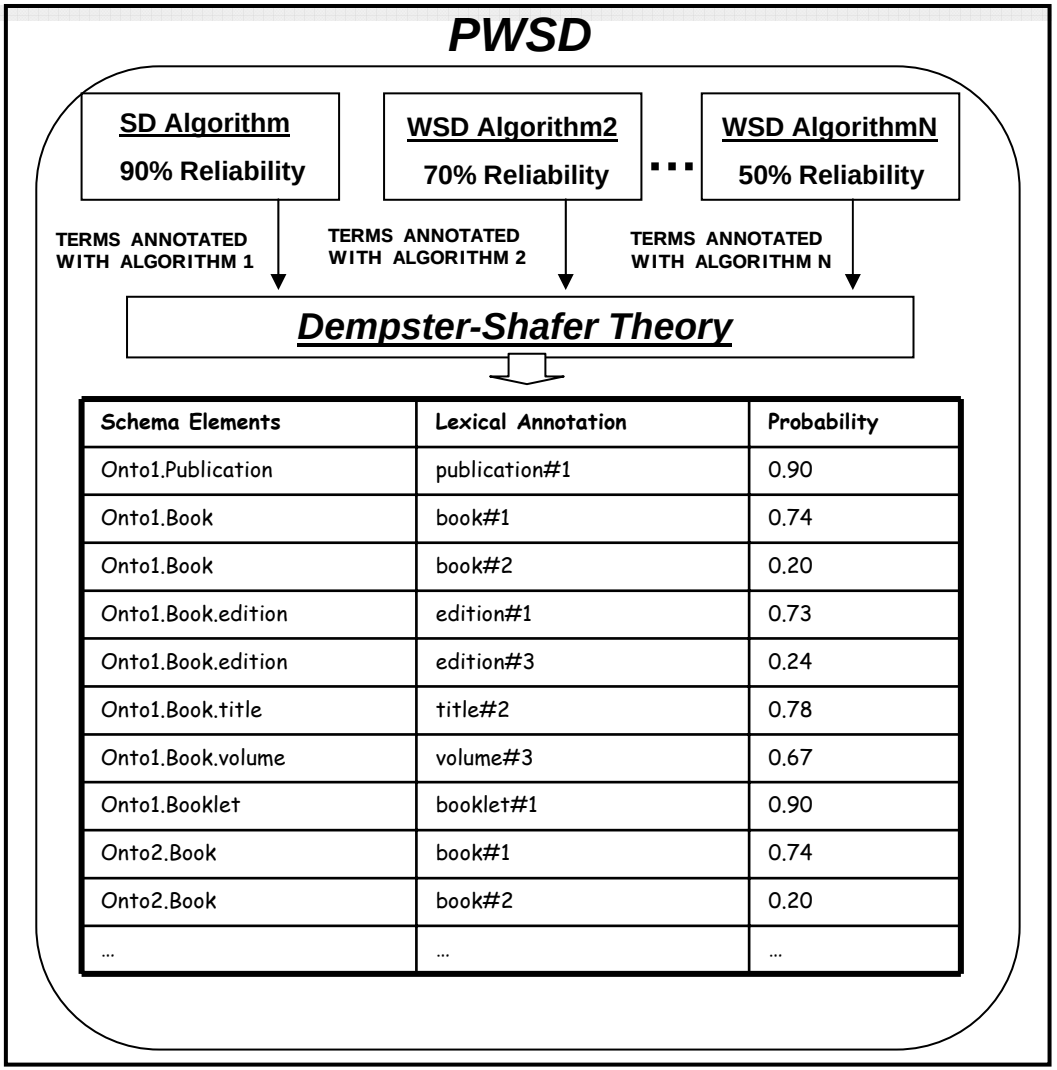
Example of application of PWSD (1)

By the use of PWSD, an automatic lexical annotation is performed over all the source elements



Example of application of PWSD (2)

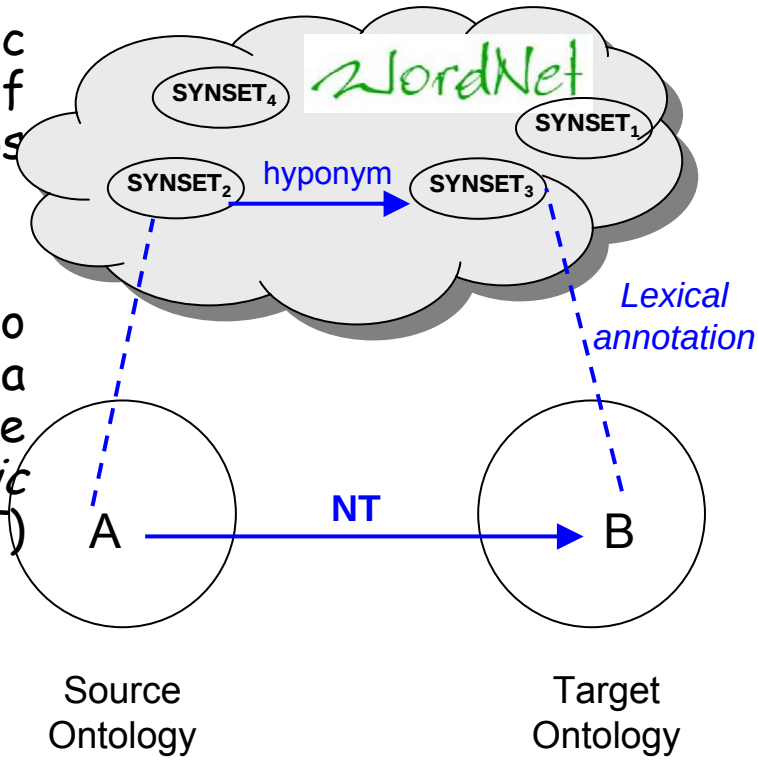
At the end of the annotation process, each source element has been associated to one or more probabilistic annotations



Schema Elements	Lexical Annotation	Probability
Onto1.Publication	publication#1	0.90
Onto1.Book	book#1	0.74
Onto1.Book	book#2	0.20
Onto1.Book.edition	edition#1	0.73
Onto1.Book.edition	edition#3	0.24
Onto1.Book.title	title#2	0.78
Onto1.Book.volume	volume#3	0.67
Onto1.Booklet	booklet#1	0.90
Onto2.Book	book#1	0.74
Onto2.Book	book#2	0.20
…	…	…

Lexical relationships extraction

- Starting from the probabilistic annotations, we derive a set of probabilistic lexical relationships between elements
- Once we have assigned meanings to source elements and discovered that a WordNet relationship hold between the meanings, we derived *probabilistic lexical relationships* (SYN, BT, NT, RT) between source elements



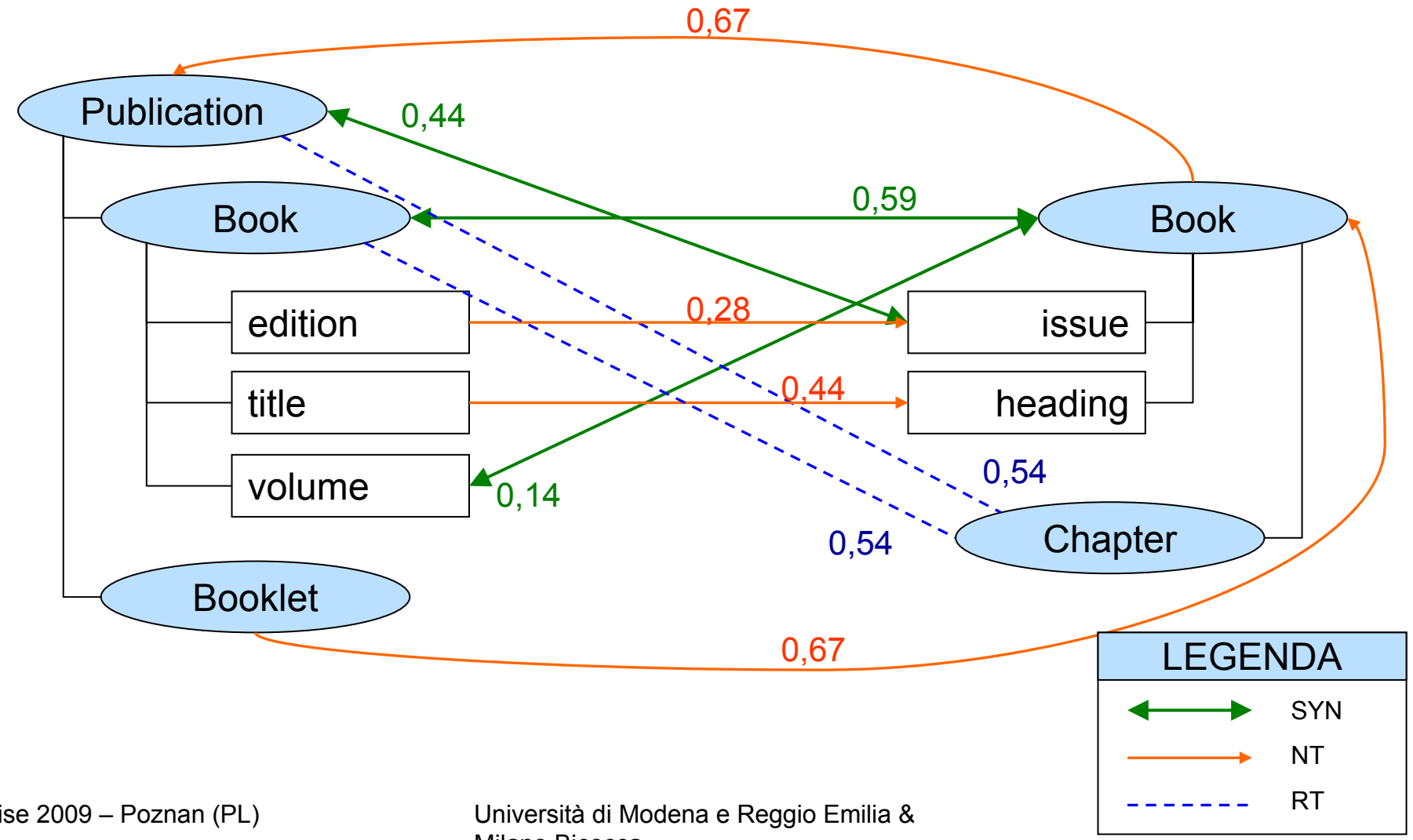
Example of lexical relationships extraction (1)

- From the lexical annotations, lexical relationships are derived
- The reliability assigned to the lexical relationships between two terms is the product of the probability value of the annotations under consideration for a term

  P(rel A,B)= P(a#i) x P(b#j)

Lexical Relationships	Probability
Onto1.Publication SYN Onto2.Book.issue	0.44
Onto1.Book SYN Onto2.Book	0.59
Onto1.Book.edition NT Onto2.Book.issue	0.28
Onto1.Book.title NT Onto2.Book.heading	0.44
Onto1.Book.volume SYN Onto2.Book	0.14
Onto1.Booklet NT Onto2.Book	0.67
Onto2.Book NT Onto1.Publication	0.67
Onto2.Chapter RT Onto1.Book	0.54
Onto2.Chapter RT Onto1.Publication	0.54

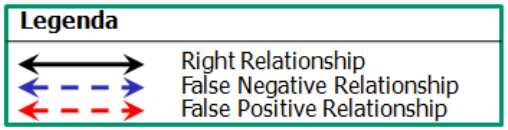
Example of lexical knowledge extraction (2)

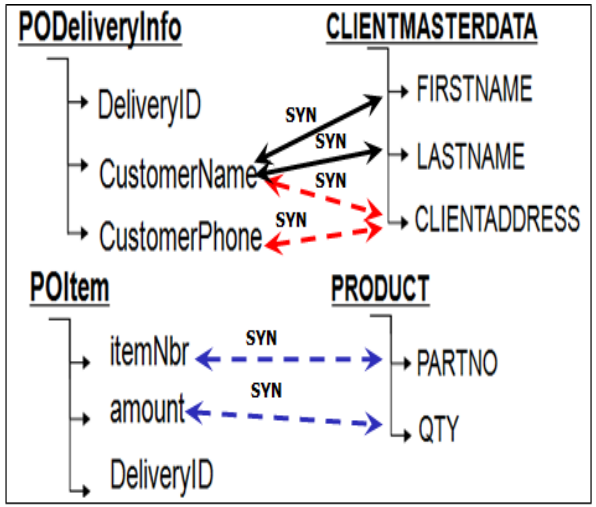


Schema labels Normalization

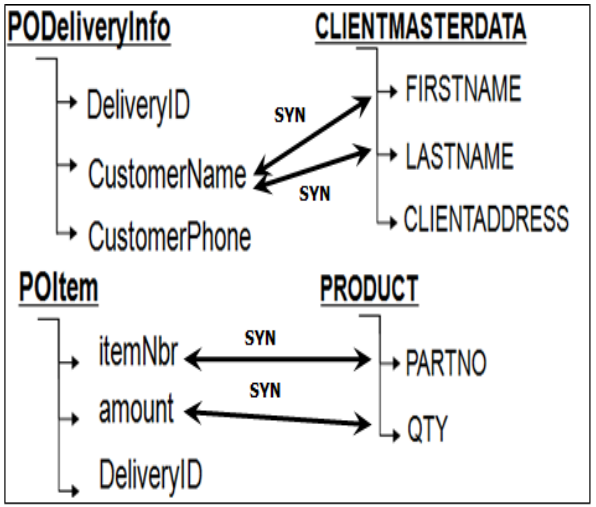
- The result of lexical annotation is strongly affected by the presence of these non-dictionary words in the schema (compound nouns, acronyms, abbreviations etc.) *i.e.* not be present in the lexical resource WordNet.
- **Schema Labels Normalization**: the reduction of the form of each label to some standardized form that can be easily recognized: in our case the process of abbreviations expansion and CNs annotation

- Improves the schema matching process by reducing the number of **false positive/false negative lexical relationships**





a- Discovered relationships without Schema Normalization



b- Discovered relationship swith Schema Normalization

Sorrentino, Bergamaschi, Gawinecki, Po – **Schema Normalization for Improving Schema Matching-** ER' 09