



Reducing terminological ambiguity: Towards standardized measures for Semantic Distance

Vipul Kashyap, National Library of Medicine, NIH

kashyap@nlm.nih.gov

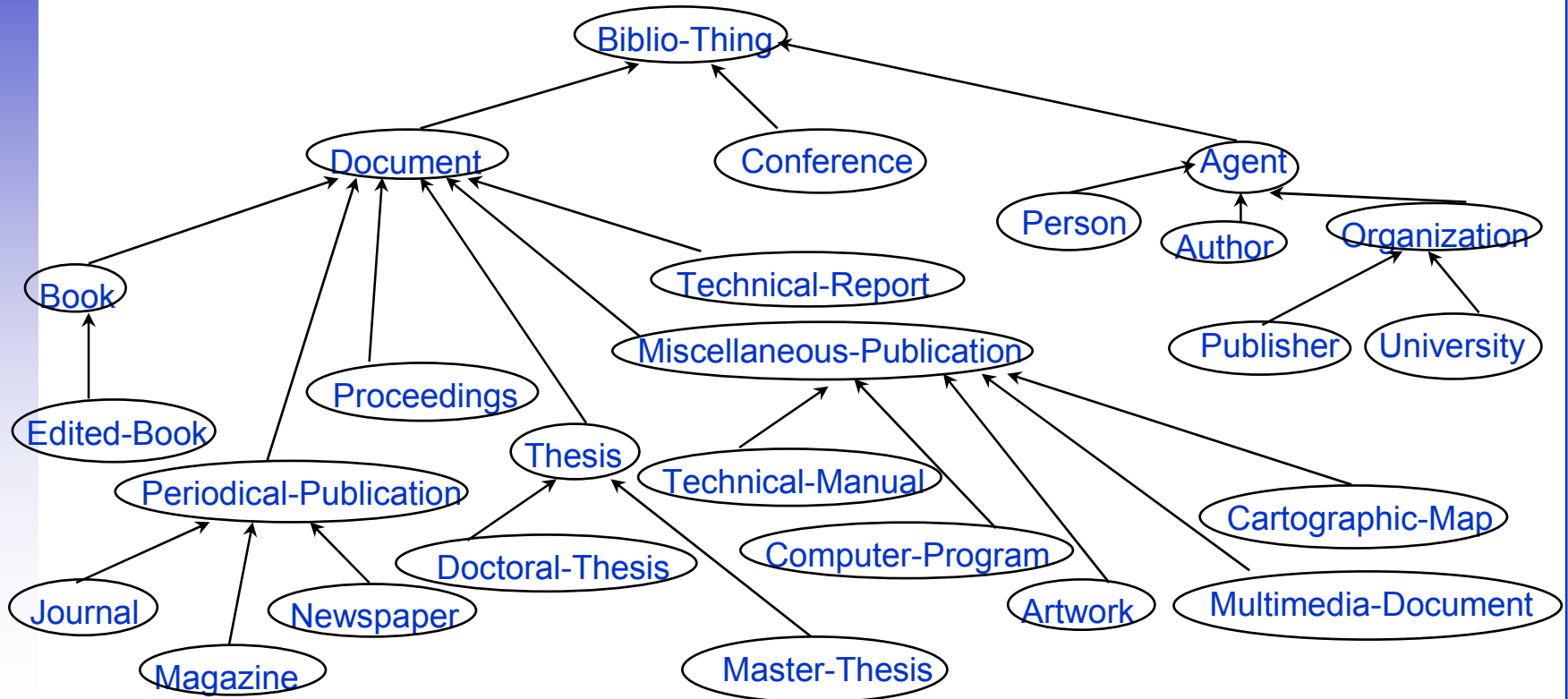
Information Technologies for Healthcare: Barriers to Implementation

NIST, Gaithersburg, MD, August 1, 2002

Motivation

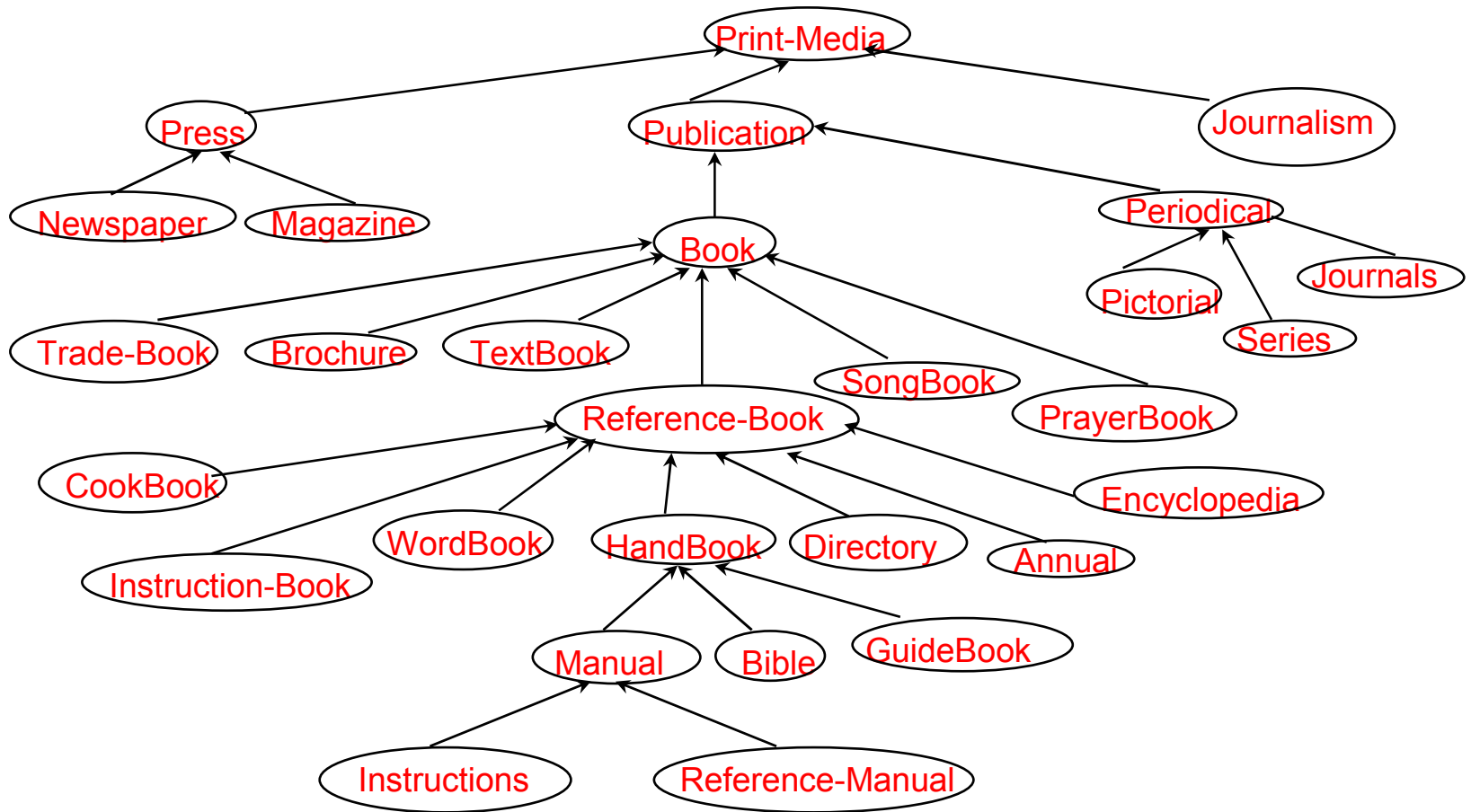
- Healthcare Information is characterized by multiple terminologies,
 - E.g., MeSH, CPT, LOINC, SnoMed, etc.
- Interoperability across terminologies is crucial to healthcare information system interoperability
- Which terminology do I interoperate with?
- What criteria/measure do we use?
 - Application dependent v/s application specific
- Should the measure be machine understandable?
- Should the measure be human understandable?

Terminology 1: The Blue Terminology



<http://www-ksl.stanford.edu/knowledge-sharing/ontologies/html/bibliographic-data/>

Terminology 2: The Red Terminology



<http://www.cogsci.princeton.edu/~wn/w3wn.html>

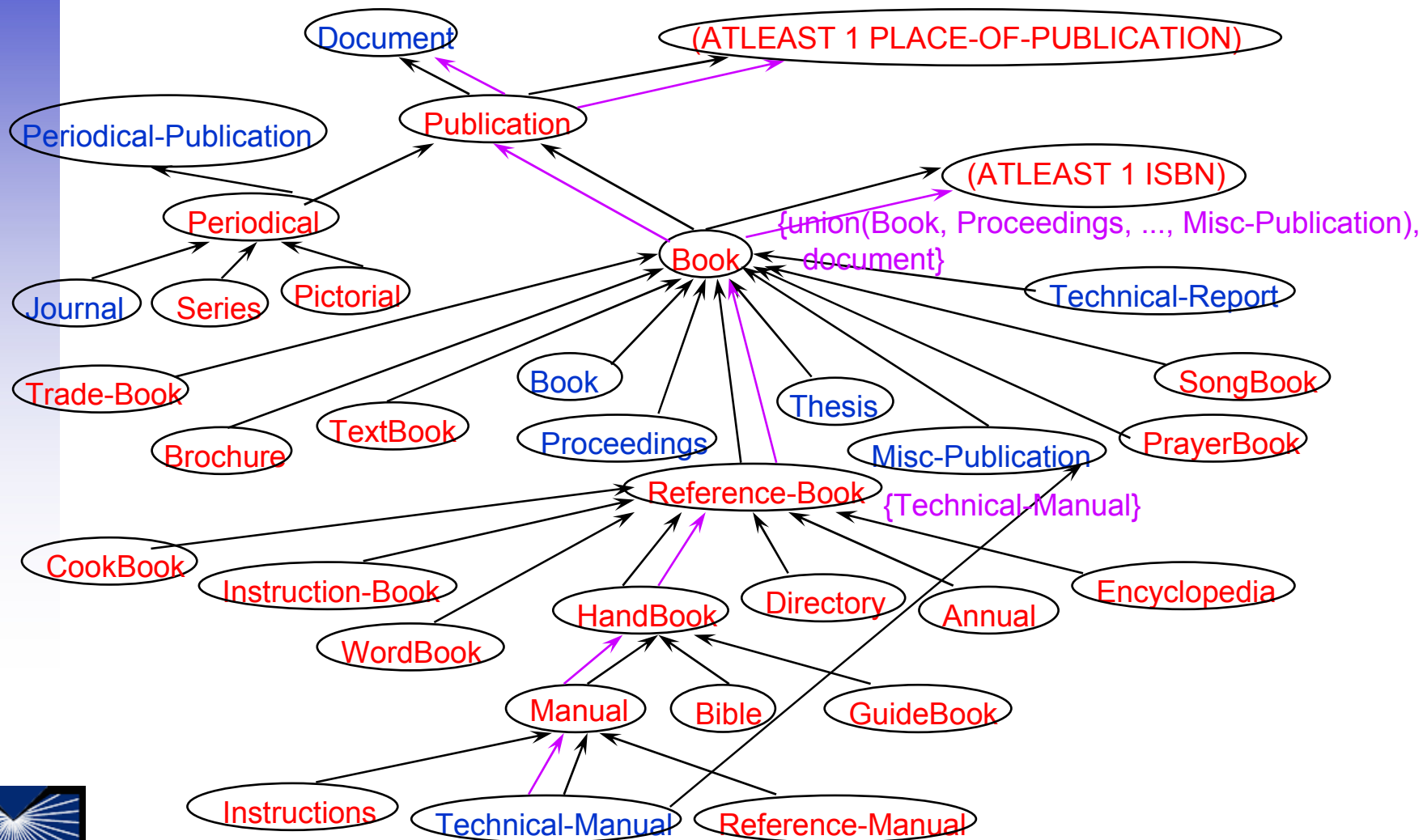
Inter-terminological relationships: Typically represented in the UMLS Metathesaurus

- Synonyms
 - semantics preserving
- Hyponyms/Hypernyms
 - semantics altering
 - typically results in loss of information

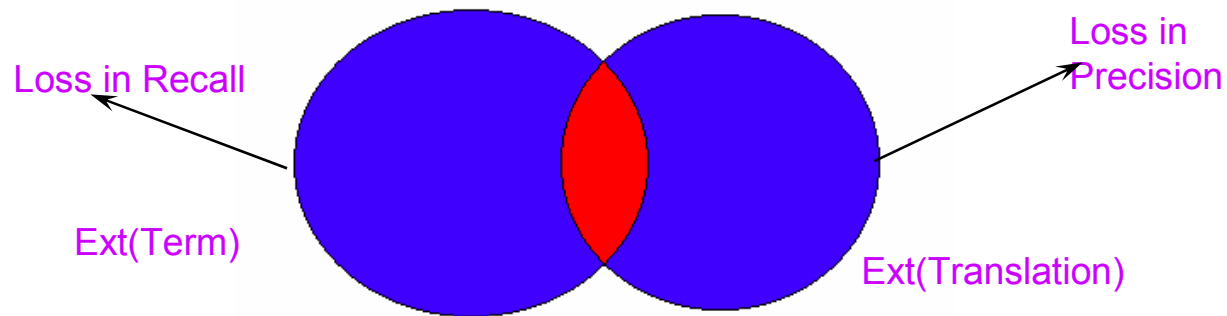
- List of Hyponyms

– technical-manual	<i>hyponym</i>	manual
– book	<i>hyponym</i>	book
– proceedings	<i>hyponym</i>	book
– thesis	<i>hyponym</i>	book
– misc-publication	<i>hyponym</i>	book
– technical-reports	<i>hyponym</i>	book
– press	<i>hyponym</i>	periodical-publication
– periodical	<i>hyponym</i>	periodical-publication

Translations across multiple terminologies



Proposal for Semantic Distance: Extensional Measure



$$\text{Precision} = \frac{|\text{Ext(Term)} \cap \text{Ext(Translation)}|}{|\text{Ext(Translation)}|}$$

$$\text{Recall} = \frac{|\text{Ext(Term)} \cap \text{Ext(Translation)}|}{|\text{Ext(Term)}|}$$

$$\text{Percentage Loss} = \frac{|\text{Ext(Term)} \Delta \text{Ext(Translation)}|}{|\text{Ext(Term)}| + |\text{Ext(Translation)}|}$$

$$= 1 - \frac{1}{1/2(1/\text{Precision}) + 1/2(1/\text{Recall})}$$

$$\Rightarrow 1 - \frac{1}{(\alpha)(1/\text{Precision}) + (1-\alpha)(1/\text{Recall})} \quad 0 < \alpha < 1$$

Using Subsumption for tighter bounds on Semantic Distance

- Term subsumes Translation
 - $\text{Ext}(\text{Translation}) \subseteq \text{Ext}(\text{Term})$
 $\Rightarrow \text{Ext}(\text{Term}) \cap \text{Ext}(\text{Translation}) = \text{Ext}(\text{Translation})$
 - Precision = 1,
 - Recall = $\frac{|\text{Ext}(\text{Translation})|}{|\text{Ext}(\text{Term})|}$
- Should be able incorporate other “application-specific” measures to adapt distance measures
- Same terminological translation might be have different semantic distances based on application specific adaptations...

Proposal for Semantic Distance: Intensional Measure

- Difference in Translation:
 - **Book** \Rightarrow union(**Book**, **Thesis**, **Proceedings**, **Technical-Manual**, **Misc-Publication**)
- Terminological Difference
 - **Book** \subseteq (AND **Publication** (ATLEAST 1 ISBN))
 - **Publication** \subseteq (AND **document** (ATLEAST 1 PLACE-OF-PUBLICATION))
 - **Book** \subseteq (AND **document** (ATLEAST 1 ISBN) (ATLEAST 1 PLACE-OF-PUBLICATION))
- Loss of Information:
 - (-) union(**Trade-Book**, **Brochure**, **SongBook**, **PrayerBook**, **TextBook**)
 - information related to trade books, brochures, song books, prayer books and text books is lost
 - (+) (AND (ATLEAST 1 ISBN) (ATLEAST 1 PLACE-OF-PUBLICATION))
 - spurious documents that don't have an ISBN number and a place of publication are gained

Measures for Semantic Distance: Pros and Cons

- Intensional Measure:
 - May not make sense as it mixes two vocabularies,
 - e.g., does **Book** - **Book** make any sense ?
 - The problem becomes worse if the two terminologies are in different languages
 - Makes it hard for the system to differentiate between the various alternatives
- Extensional Measure:
 - Based on Standard Information Retrieval Measures (F-measure)
 - Can be tailored to reflect change in semantic distance for different applications
 - However:
 - Probability distributions of various terms need to be estimated
 - An information loss interval doesn't make much sense to the user.

Conclusions

- Semantic Distance measures need to be application specific:
 - Text Retrieval
 - (Structured) Data Retrieval
 - Domain and Context Specific
- Semantic Distance measures should be both human and machine processable
- They should be based on standard measures as far as possible
 - E.g., F-measure from Information Retrieval
- There is a need for estimation of various distributions of medical concepts in a given population:
 - E.g. May need to mine CDC databases

Proposal for Semantic Distance: Tversky's measure from Psycho-semantics

$$S(a, b) = \frac{|A \cap B|}{|A \cap B| + \alpha(a, b) |A - B| + (1 - \alpha(a, b)) |B - A|}$$

- $S(a, b)$ is the similarity between two arbitrary objects, a, b
- A and B are feature sets of a, b respectively
- α is a real no. $\exists 0 \leq \alpha \leq 1$