USC

USC Viterbi School of Engineering

Determining Semantic Similarity among Entity Classes from Different Ontologies

Authors: M. Andrea Rodriguez, Max J. Egenhofer

Presented By: Yaoxuan Wang

Overview

- Introduction
- Approach
- Experiments
- Conclusion
- Discussion

- **Motivation**
  - Need new tools that can improve the retrieval and integration of information

- Previous
  - Compare concepts from different ontologies are based on an a priori integration of local ontologies

- This work
  - Create a computational model to assess semantic similarity among entity classes from unconnected and independent ontologies
  - No integration of ontologies is needed

- Entity Class:
  - refers to concepts that group entities or objects of the real world into classes
  - e.g. building, lake , city, etc.
- Entity Class Representation:
  - **Synonym set**
    - a set of synonym words that denotes an entity class
    - address polysemy and synonymy
    - e.g. bank, depository financial institution
  - **Semantic interrelation**
    - determine general organization of entity classes
    - Hyponymy, i.e. “is-a” relation equal to “inheritance”
    - Meronymy, i.e. “part-whole” relation equal to “composition”
  - **Distinguishing features**
    - distinguish entity classes from the same superclass besides semantic interrelation
    - e.g. hospital vs. apartment building
  - **Classification**
    - functions, what is done to or with instances of a class, e.g. practice of stadium
    - parts, structural elements of a class, e.g. roof, floor of a building
    - attributes, the rest

- Recall:
  - the fraction of similar entity classes that are detected by the model

$$recall = \frac{|A \cap B|}{|A|}, \qquad (13a)$$

- **Precision**:
  - the fraction of entity classes detected by the model that are actually similar

$$precision = \frac{|A \cap B|}{|B|}. \qquad (13b)$$

- **A** is the set of similar entity classes
- **B** is the set of similar entity classes calculated by the model
- | | is the counting measure or cardinality

TABLE 1
Entity_Class Definition in BNF Notation, and an Example of the Definition of *Stadium*

BNF Notation	Example: *Stadium*
<entity_class>::= **entity_class {** **name:** {<syn_set>} **description:** <description> **is_a:** <is-a> **part_of:** <part_of> **whole_of:** <whole_of> **parts:** <parts> **functions:** <functions> **attributes:** <attributes>**}**	**entity_class {** **name:** {stadium,bowl,arena} **description:** large often unroofed structure in which athletic events are held **is_a:** {*construction**} **part_of:** { } **whole_of:** {*athletic_field**} **parts:** {{athletic_field,sports_field,playing_field}, {dressing_room},{foundation}, {midfield},{spectator_stands,stands}, {ticket_office, box_office,ticket_booth}} **functions:** {{play,compete},{play,practise}, {recreate,play}} **attributes:** {{architectural_property}, {covered/uncovered}, {name}, {lighted/unlighted},{owner_type}, {sports_type},{user_type}}}
<is_a>::= { }|{<pts_entity_classes>}	
<part_of>::= { }|{< pts_entity_classes >}	
<whole_of>::= { }{< pts _entity_classes >}	
<parts>::= { }|{<syn_sets>}	
<functions>::= { }|{<syn_sets>}	
<attributes>::= { }|{<syn_sets>}	
<syn_sets>::= {<syn_set>}|<syn_sets>,{syn_set}	
<syn_set>::= <word>|<syn_set>,<word>	
<description>::= <word>|<description> <word>	
< pts _entity_classes>::= <pointer>| <pt_to_entity_classes>,<pointer>	

(x^* denotes a pointer to the entity class x)

- **Entity Class Representation Comparison**
  - the representation of entity classes across ontologies should share some components
- **Three independent similarity assessments**
  - **Synonym sets**
    - synonyms refer to the same entity class
    - the similarity between an entity class and itself is maximal
    - exploit the general agreement in the use of words
    - detect equivalent words that likely refer to the same entity class
    - BUT only a basic level of similarity assessment
    - e.g. clinic vs. hospital, polysemy
  - **Distinguishing feature**
    - determine how similar entity classes are
    - Only have some common feature, e.g. stadium vs. sports_arena
  - **Semantic relations**
    - whether target entity classes are related to the same set of entity classes
    - e.g. hospital vs. house – building
    - how to compare?
      - a comparison between the **semantic neighborhoods** of entity classes

- **Semantic neighborhood (N)**

$$N(a^o, r) = \{c_i^o\} \text{ such that } \forall i \; d(a^o, c_i^o) \leq r. \qquad (1)$$

  - a^o : the entity class
  - c_i^o : the set of entity classes whose distance d() <= r
  - r : nonnegative integer
  - d(): the shortest path in the ontology between two entity classes, i.e. the smallest number of arcs

  - the semantic neighborhood of an entity class also contains itself

- **Similarity of semantic relation**

  - NOT based on this path distance

  - Path distance => neighborhoods
  - the similarity between entity class A and B depends on the similarity of the entity classes in their neighborhoods

$$N(a^o, r) = \{c_i^o\} \text{ such that } \forall i \; d(a^o, c_i^o) \leq r. \qquad (1)$$

- a^o = stadium in WordNet ontology
- r = 1

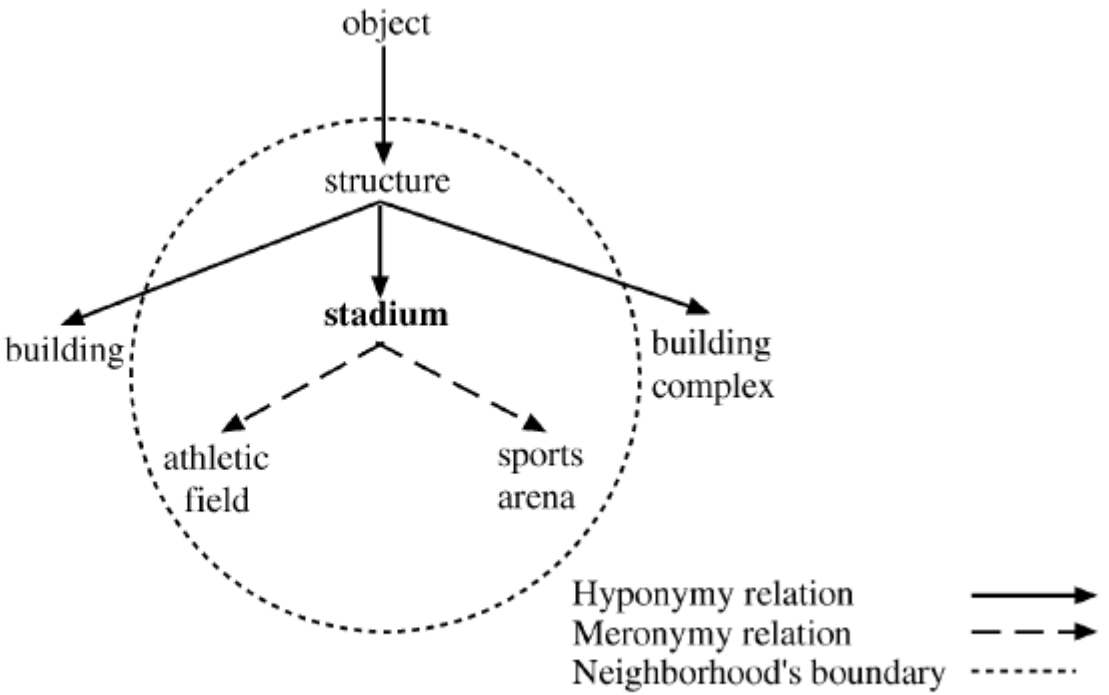

Fig. 1. Example of the immediate semantic neighborhood of *stadium* in a portion of the WordNet ontology.

- **Integrate the similarity assessments**

$$S(a^p, a^q) = \omega_w \cdot S_w(a^p, b^q) + \omega_u \cdot S_u(a^p, b^q) + \omega_n \cdot S_n(a^p, b^q),$$
$$for\ \omega_w, \omega_u, \omega_n \geq 0. \qquad (2)$$

- S(): the similarity between entity class *a* in ontology *p* and *b* in ontology *q.*
- Sw: the similarity of synonym sets
- Su: the similarity of features
- Sn: the similarity of semantic neighborhoods
- ωw: the weight of the similarity of synonym sets component
- ωu: the weight of the similarity of features component
- ωn: the weight of the similarity of semantic neighborhoods component

- **A Matching Model to Similarity Assessment**

$$S(a,b) = \frac{|A \cap B|}{|A \cap B| + \alpha(a,b)|A/B| + (1-\alpha(a,b))|B/A|}, \quad (3)$$

for $0 \leq \alpha \leq 1$.

- S(a, b): the similarity of entity classes *a* and *b* calculated by matching model
- A, B : description sets of *a* and *b*, i.e. synonym sets, sets of distinguishing features, and sets of entity classes in the semantic neighborhood
- A∩B : intersection of A and B
- A / B : difference of A and B
- | | : the cardinality of a set
- α(a, b) : defines the relative importance of the noncommon characteristics
- a : referred to as the **target**
- b : referred to as the **base**

- the matching model is not forced to satisfy metric properties
  - e.g. S(building, office_building) != S(office_building, building)

- **A Matching Model (2)**

$$\alpha(a^p, b^q) = \begin{cases} \frac{depth(a^p)}{depth(a^p)+depth(b^q)} & depth(a^p) \leq depth(b^q) \\ 1 - \frac{depth(a^p)}{depth(a^p)+depth(b^q)} & depth(a^p) > depth(b^q). \end{cases} \quad (4)$$

- In a cross-ontology, two independent ontologies are connected by making each of their roots a direct descendant of an imaginary and more general entity class, called *anything.*

- depth() : the shortest path from the entity class to the imaginary root
  - reflects the degree of granularity upon which the ontology was designed

- value of α are greater than 0 and less than or equal to 0.5

- **A Matching Approach (3)**

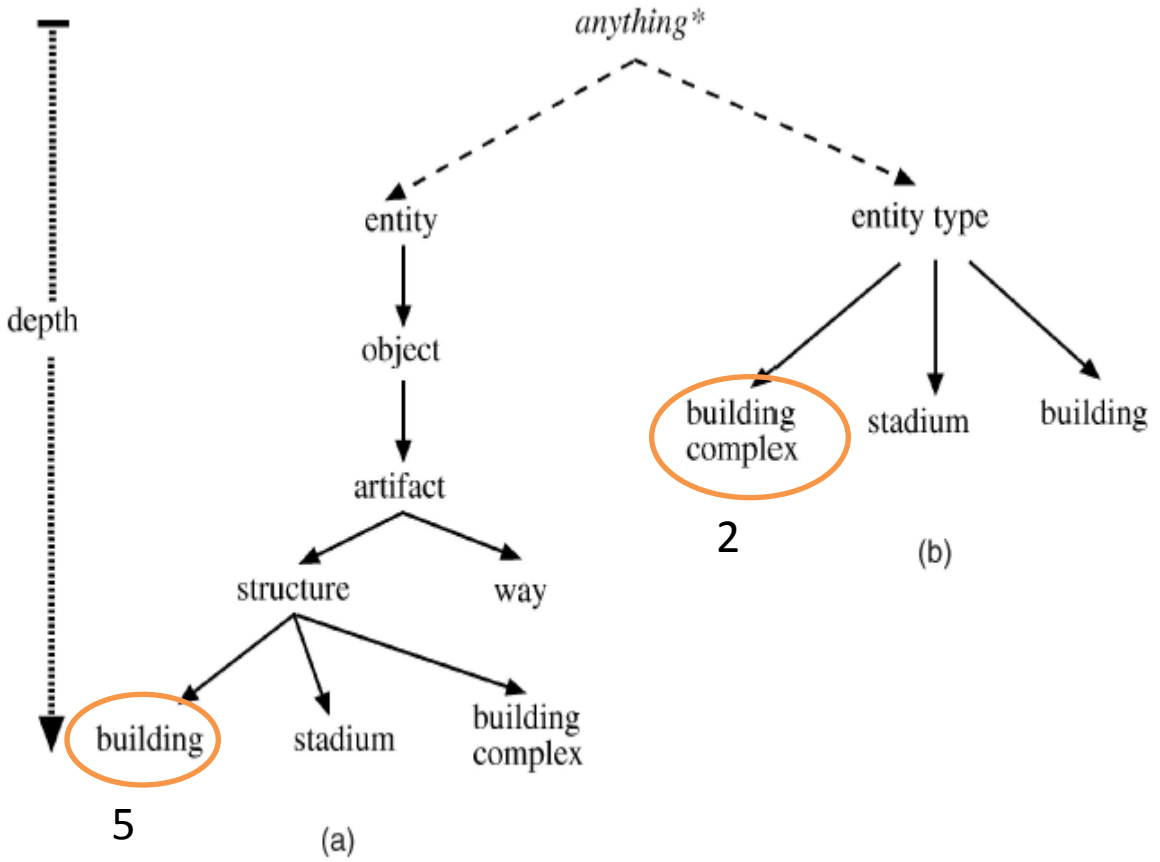

Fig. 2. Connecting independent ontologies: (a) partial WordNet ontology and (b) partial SDTS ontology. (*Anything** denotes an imaginary root.)

- **Word Matching**

$$S_l(building^w, building\_complex^s) = \frac{|\{building\}|}{|\{building\}| + 0.28|\{\}| + 0.72|\{complex\}|} = \frac{1}{1.72} = 0.58. \quad (5)$$

2/(5+2)

- **Feature Matching**

$$S_u(a^p, b^q) = \omega_p \cdot S_p(a^p, b^q) + \omega_f \cdot S_f(a^p, b^q) + \omega_a \cdot S_a(a^p, b^q)$$
$$\text{for } \omega_p, \omega_f, \text{and } \omega_a \geq 0 \text{ and } \omega_p + \omega_f + \omega_a + 1.0. \quad (6)$$

- Su() : the similarity between entity class *a* in ontology *p* and *b* in ontology *q.*
- Sp, Sf, Sa : the similarity of parts, functions, attributes
- ωp, ωf, ωa : the weight of the similarity of each specification component

- represented by a synonym set
- **strict string matching**, only if represented by the same word or by synonym sets that intersect

- two distinguishing features are equivalent if the intersection of their synonym sets is not empty

TABLE 2
Entity_Class Definition of *Stadium* in WS and WordNet

Stadium (WS)	**Stadium (WordNet)**
entity_class {	**entity_class {**
name: {stadium,bowl,arena}	**name:** {stadium,bowl,arena}
description: large often unroofed structure in which athletic events are held	**description:** large often unroofed structure in which athletic events are held
is_a: {*construction**}	**is_a:** {*construction**}
part_of: {}	**part_of:** {}
whole_of: {*athletic_field**}	**whole_of:** {*athletic_field**, *sports_arena**}
parts: {{athletic_field,sports_field,playing_field}, {dressing_room},{foundation}, {midfield},{spectator_stands,stands}, {ticket_office, box_office,ticket_booth}}	**parts:** {{athletic_field,sports_field,playing_field}, {foundation},{midfield},{plate}, {sports_arena,field_house},{stands}, {structural_elements}, {standing_room},{tiered_seats}}
functions: {{play,compete},{play,practise}, {recreate,play}}	**functions:** {}
attributes: {{architectural_property}, {covered/uncovered}, {name}, {lighted/unlighted},{owner_type}, {sports_type},{user_type}}}	**attributes:** {}}

(x denotes a pointer to the entity class x)*

$$X = stadium^{ws}.\text{parts} \cap stadium^{w}.\text{parts}$$
$$= \{\{athletic\_field, playing\_field, field\}, \{foundation\}, \{midfield\}, \{stands\}\}. \quad (7)$$

$$Y = stadium^{w}.\text{parts} - stadium^{ws}.\text{parts}$$
$$= \{\{plate\}, \{sports\_area, field\_house\}, \{standing\_room\}, \{structural\_elements\}, \{tiered\_seats\}\} \quad (8a)$$

$$Z = stadium^{ws}.\text{parts} - stadium^{w}.\text{parts}$$
$$= \{\{dressing\_room\}, \{ticket\_office, box\_office, ticket\_booth\}\}. \quad (8b)$$

$$S_u(stadium^{w}, stadium^{ws}) = S_p(stadium^{w}, stadium^{ws})$$
$$= \frac{|X|}{|X| + 0.45|Y| + 0.55|Z|}$$
$$= \frac{4}{4 + 0.45 * 5 + 0.55 * 2} = 0.54. \quad (9)$$

- **Semantic-Neighborhood Matching**
  - compare entity classes in semantic neighborhoods based on **word** or **feature matching**

$$S_n(a^p, b^q, r) = \frac{|a^p \cap_n b^q|}{|a^p \cap_n b^q| + \alpha(a^p, b^q) \cdot \delta(a^p, a^p \cap_n b^q, r) + (1 - \alpha(a^p, b^q)) \cdot \delta(b^q, a^p \cap_n b^q, r)}$$

$$\text{with } \delta(a^p, a^p \cap_n b^q, r) = \begin{cases} |N(a^p, r)| - |a^p \cap_n b^q| & \text{if } |N(a^p, r)| > |a^p \cap_n b^q| \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

- **Numerator**
  - $|a^p \cap_n b^q|$ is the approximate cardinality of the set intersection between these semantic neighborhoods
  - a_i^p and b_j^q are entity classes in the semantic neighborhood of a^p and b^q
  - n and m are the numbers of entity classes in the corresponding semantic neighborhoods

$$|a^p \cap_n b^q| = \left[\sum_{i \leq n} max_{j \leq m}\ S(a_i^p, b_j^q)\right] - \varphi S(a^p, b^q), \text{ where}$$

$$\varphi = \begin{cases} 1 & \text{if } S(a^p, b^q) = max_{j \leq m}\ S(a^p, b_j^q) \\ & \text{and} \\ 0 & \text{otherwise} \end{cases}$$

$$s(a_i^p, b_j^q) = \omega_l' S_l(a_i^p, b_j^q) + \omega_u' S_u(a_i^p, b_j^q) \text{ with } 0 < \omega_l' + \omega_u' \leq 1. \quad (11)$$

- **Cross-Ontology Evaluation**
  - Address the quality of results of similarity assessments
  - Design new experiments
    - have this matching model for similarity evaluations
    - use available ontologies, WordNet and SDTS
    - use human-subject testing
  - Create a new ontology WS from the combination of WordNet and SDTS to exploit a more complete definition of entity classes
  - **Two type of experiments**
    - searching for equivalent or most similar entity classes across ontologies
      - useful for ontology integration
    - ranking the similarity between entity classes in two ontologies
      - useful for information retrieval
      - e.g. stadium and athletic field

TABLE 3
Characteristics of the Specification Componenets of SDTS, WordNet, and WS

Characteristics	SDTS	WordNet	WS
Words			
Synonymy		√	√
Polysemy	√	√	√
Relations			
Is-a	√	√	√
Part-of		√	√
Whole-of		√	√
Features			
Parts		√	√
Functions			√
Attributes	√		√

Experiments (1)

- Combinations:
  - identical ontologies (1-2)
  - ontology and sub ontology (3)
  - overlapping ontologies (4)
  - different ontologies (5)
- Assessment measures:
  - use **recall** and **precision**
  - what entity classes are in fact similar?
    - equivalent or most similar entity classes, e.g. building vs. (building, building complex)

TABLE 4
Cases of Cross-Ontology Evaluations

Case	Ontology-Ontology	Description
1	WordNet-WordNet	Same ontology with is-a and part-whole relations
2	SDTS-SDTS	Same ontology with is-a relations and attributes
3	WordNet-WordNet*	Subset with same specification components
4	WordNet*-WS	Overlapping semantic relations and attributes
5	WordNet*-SDTS*	Different ontologies and specification components

*(Symbol * denotes small subsets of the initial ontology)*

Experiments (1)

USC Viterbi School of Engineering

TABLE 5
Recall and Precision of Evaluations with Threshold Equal to 75 Percent

Case	Weights (%)			Recall	Precision
	ω_w	ω_u	ω_n	(%)	(%)
WordNet-WordNet	50	0	50	100	97
WordNet-WordNet	0	100	0	48	10
SDTS-SDTS	50	0	50	100	100
SDTS-SDTS	0	0	100	100	1
WordNet-WordNet*	50	0	50	99	98
WordNet-WordNet*	0	50	50	28	14
WordNet*-WS	100	0	0	100	78
WordNet*-WS	50	0	50	55	98
WordNet*-WS	0	50	50	0	0
WordNet*-SDTS*	100	0	0	100	42
WordNet*-SDTS*	50	0	50	50	92
WordNet*-SDTS*	0	100	0	0	0

*(Symbol * denotes small subsets of the entire ontology)*

- Conclusions:
  - ontologies share more components, the model produces more accurate results
  - the synonym sets and semantic neighborhood of entity classes is more similar across ontologies than the distinguishing features

USC

Experiments (2)

- Comparison:
  - an entity class in an ontology with a reduced set of entity classes defined in another ontology.
  - e.g. stadium vs. stadium, athletic field, ballpark, tennis court, commons, building, theater, museum, library, transportation system, house, and sport arena.
- Human-subject test:
  - ask subjects to rank the similarity among the set of entity classes based on the definitions in the WS ontology
  - assume a number of ranks equal to the number of entity classes compared
  - Subjects found that the most similar entity classes to a stadium in decreasing order were: sports arena, ball park, athletic field, tennis court, theater museum, building, commons, library, house and transportation
- Type of Weight setting:
  - ωw, ωu, ωn: (0.33, 0.33, 0.33), (0.5, 0, 0.5), (0, 100, 0)

- The ordering of entity classes along the horizontal axis corresponds to the subjects' responses in decreasing order

- SDTS-WS:

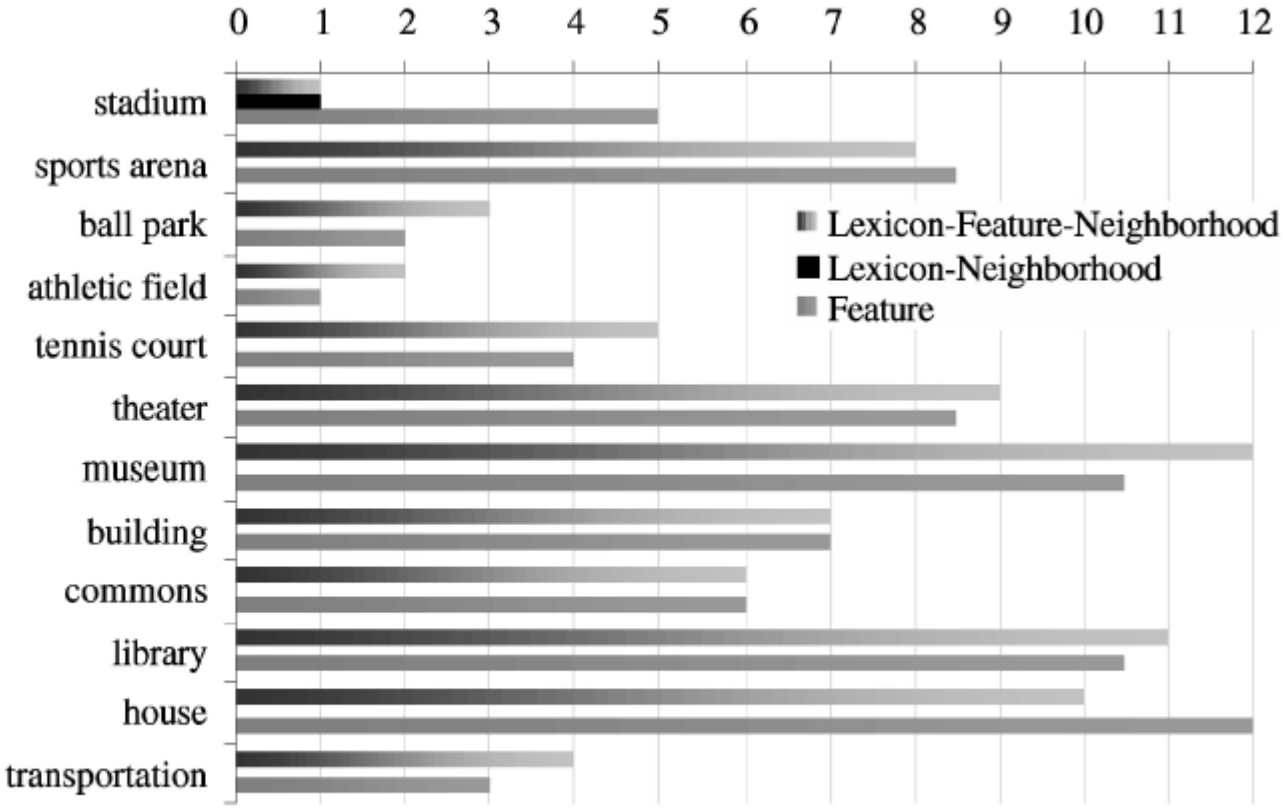

Fig. 3. Evaluations between the definition of *stadium* in SDTS and a set of entity classes defined in WS.

- WordNet-WS:

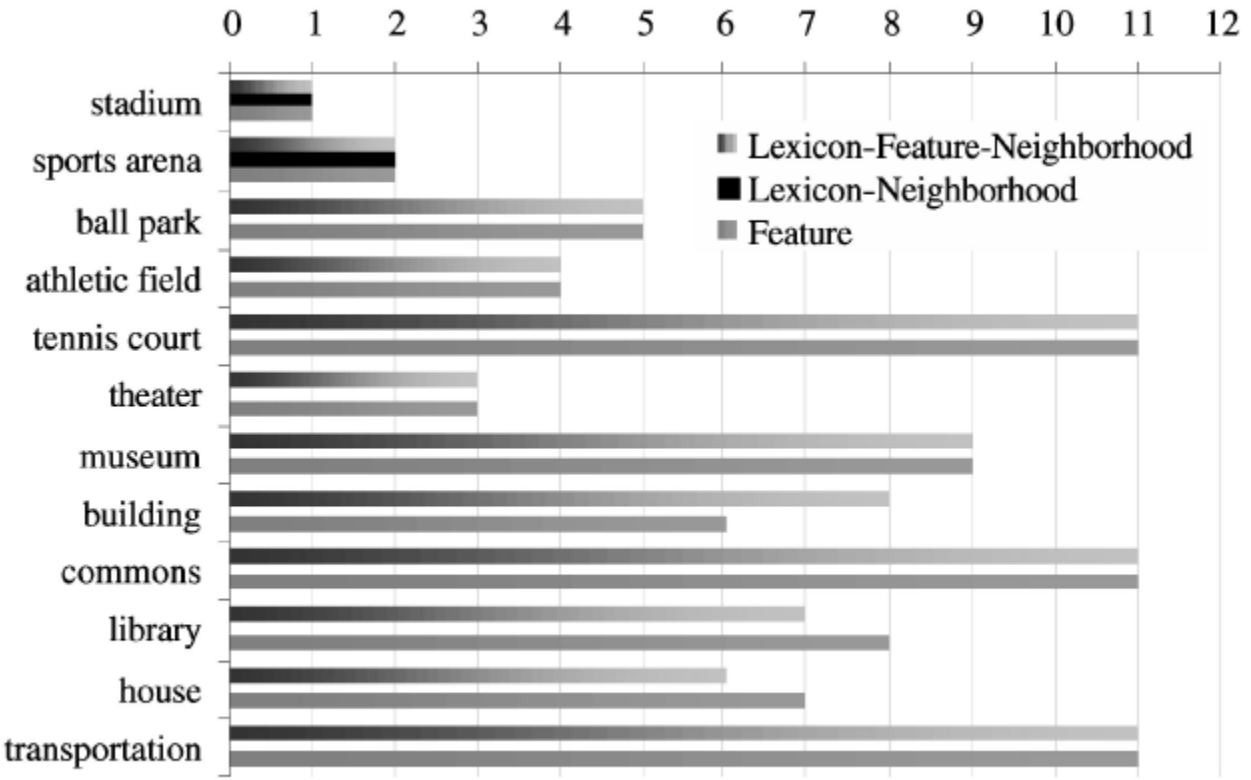

Fig. 4. Evaluations between the definition of *stadium* in WordNet and a set of entity classes defined in WS.

- WS-WS:

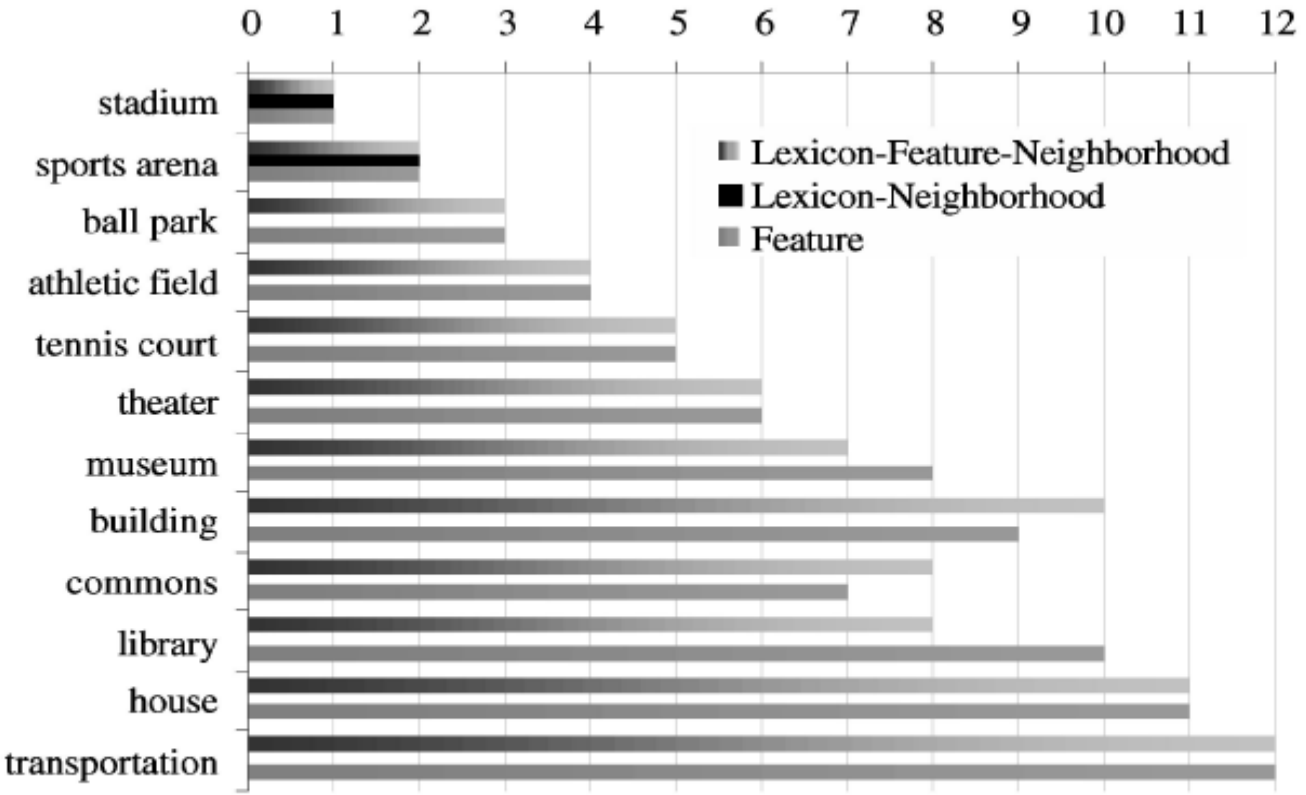

Fig. 5. Evaluations between the defintion of *stadium* in WS and a set of entity classes defined in WS.

TABLE 6
Correlation Coefficient for Similarity Ranks in Cross-Ontology Evaulations

Ontologies	Word-Feature-Neighborhood ω_w: 33.3, ω_u: 33.3, ω_n: 33.3	Word-Neighborhood ω_w: 50, ω_n: 50	Feature ω_u: 100
SDTS-WS	0.48	-0.34	0.37
WordNet-WS	0.68	-0.34	0.71
WS-WS	0.96	-0.34	0.97

correlation coefficient

- Conclusions:
  - ontologies share more components, the model produces more accurate results
  - the best combination of weights detected in the former experiment give the worst value of correlation, BECAUSE detect the most similar entity class and nothing else
  - feature matching is important for detecting similar entity classes within an ontology or the similarity of semantically related entity classes across ontologies
- Assignment of weight:
  - ontology characteristics
  - the goal of the similarity assessment
    - i.e. ontology integration vs. information retrieval

Conclusion

USC Viterbi School of Engineering

- Similarity model:
  - based on the matching process of each component
  - useful as a first step in an ontology integration

- New insight:
  - **Synonym set** and **Semantic neighborhoods**
    - for detecting equivalent or most similar entity
  - **Distinguishing features**
    - for detecting entity classes that are somewhat similar
    - i.e. not synonyms and are located far apart in the hierarchical structure
    - e.g. stadium and athletic field

Discussion

- Further work needed:
  - Now compare distinguishing feature in terms of a strict string matching between synonym sets refer to those feature, SO **semantic similarity** among features is needed.
    - Parts are also entity classes that could be semantically compared in a recursive process.
    - Verbs could be related by the semantic relation **entailment** (e.g. buy and pay)

Thank You

USC Viterbi School of Engineering

USC