Combining a co-occurrence-based and a semantic measure for entity linking

ESWC 2013: Extended Semantic Web Conference
28 May 2013, Montpellier, France

Bernardo Pereira Nunes, Stefan Dietze, Marco Antonio Casanova,
Ricardo Kawase, **Besnik Fetahu**, Wolfgang Nejdl
(PUC-Rio, BR) (L3S Research Center, DE)

Outline

- Introduction
- Motivation Example
- A combined approach towards entity linking
  - Semantic Connectivity Score – Katz Index
  - Co-occurrence-based measures
  - Combined entity linking approach
- Evaluation
- Results
- Conclusions

Introduction

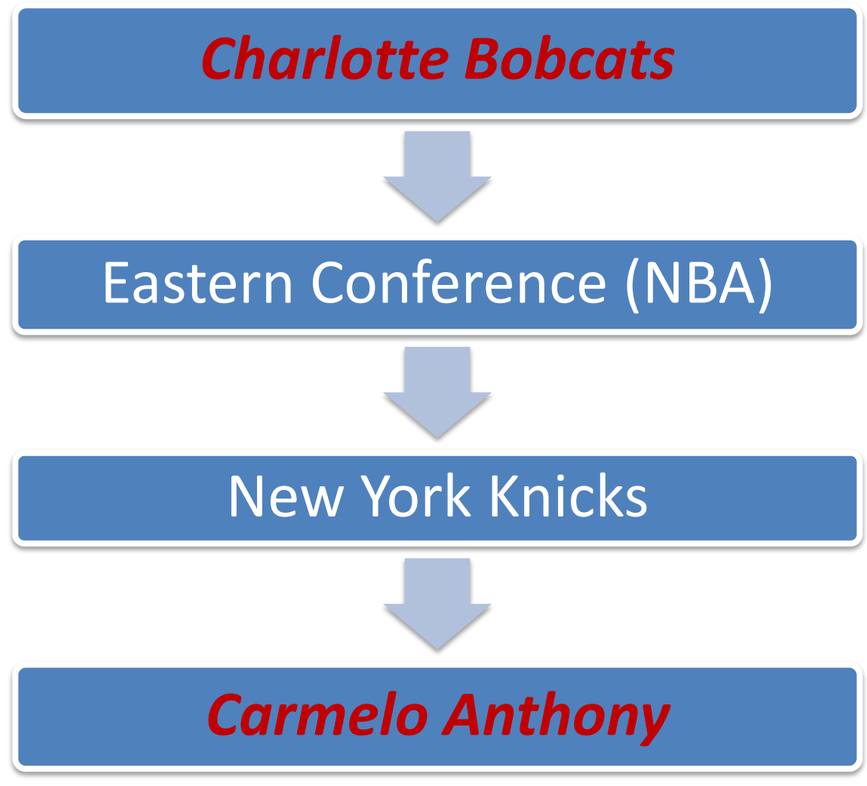
- Linked Data and Web resources
- Sparsely interlinked resources
- Knowledge bases, with structured knowledge about entities
- NER & NED for extraction of entities
- Few semantics relationships between entities (*skos:related, so:related*)
- Entity linking, meaningful only at first (direct) degree of connectivity
- Exhaustive process considering large amounts of resources

Motivation Example

- Semantic relatedness of concepts (entities)
- Exploit existing knowledge base structures
- Resource semantic similarity (entities)
- Latent relationships via semantic relations

• The **Charlotte Bobcats** could go from the **NBA**'s worst team to its best bargain.

•The **New York Knicks** got the big-game performances they desperately needed from **Carmelo Anthony** and **Amar'e Stoudemire** to beat the **Miami Heat**.



A combined entity-linking approach

Novel approach on entity-linking across resources of same and disparate datasets.

1. ***Semantic Connectivity Score*** **(SCS)–** knowledge graph based on Social Network Theory – Katz Index.
2. ***Co-occurrence based Measure*** **(CBM)** – utilise entity co-occurrence in the Web.

Semantic Connectivity Score - SCS

- Measure relatedness of entity pairs computing Katz's Index
- Use *transversal* properties to compute relatedness
- Exclude *hierarchical* properties:
  - rdfs:subClassOf
  - dcterms:subject
  - skos:broader
- Quantify semantic connectivity of entity pairs (e_1, e_2):

$$SCS(e_1, e_2) = \sum_{l=1}^{\tau} \beta^l \cdot | \, paths_{(e_1,e_2)}^{<l>} \, |$$

damping factor, exponentially penalize longer paths.

transversal paths of length *l* between entity pairs

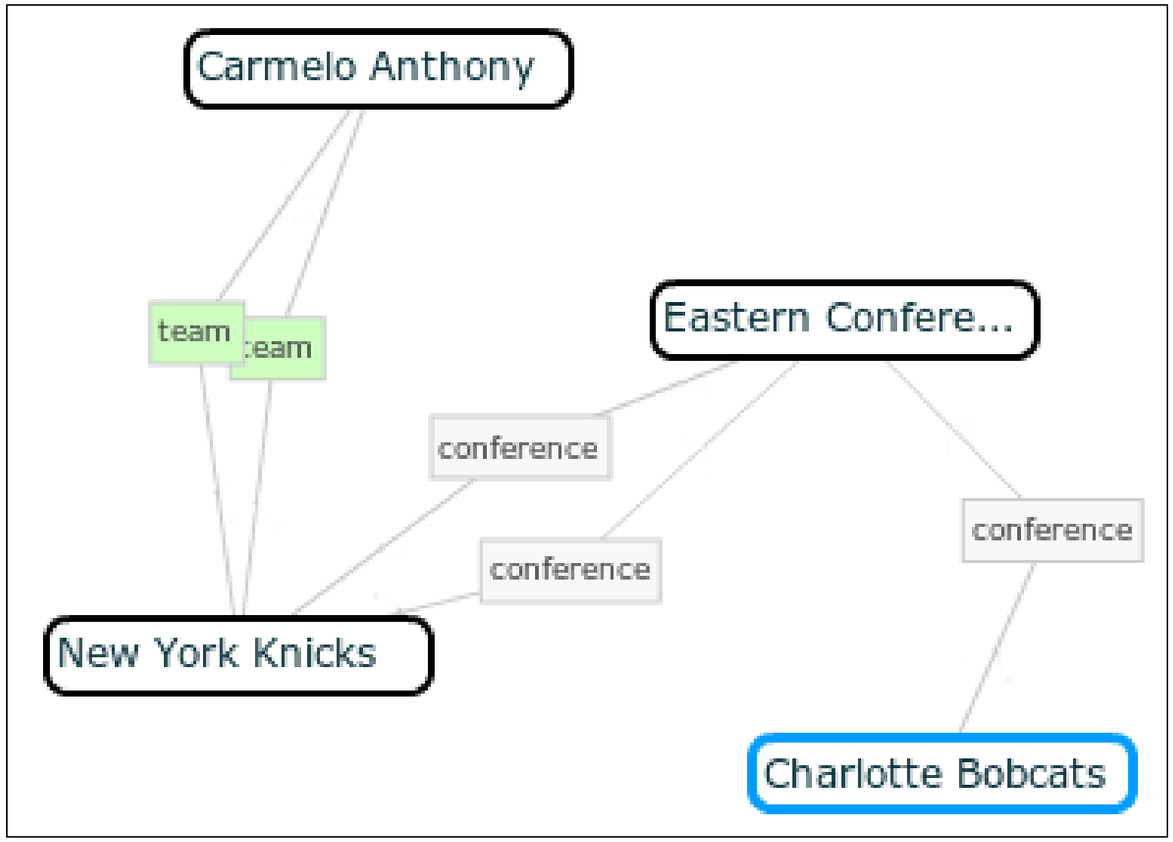
Semantic Connectivity Score – SCS (1)

Adoptions to knowledge graphs towards applying *Katz index* measure

- Remove edge directions from graphs
- Inverse properties considered equivalent:

  i.e. *isFathorOf* ↔ *isSonOf*

- Empirically determine *path length*

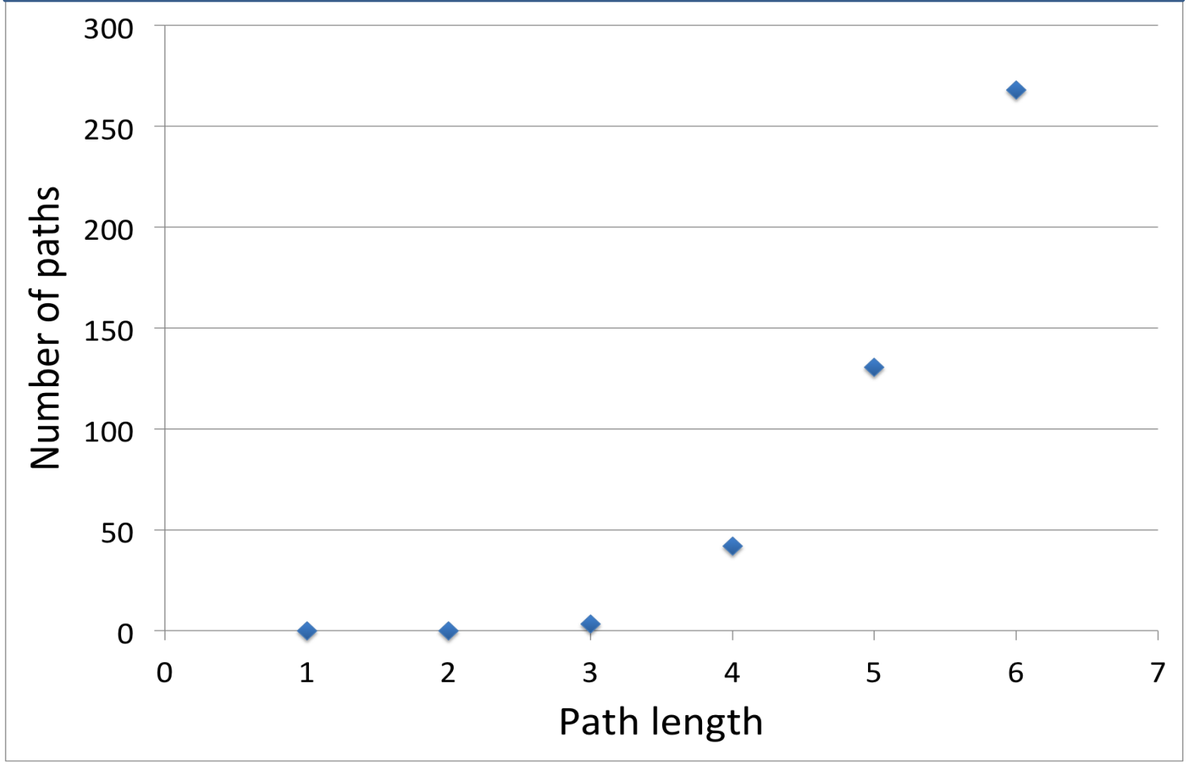


Inverse property equivalence

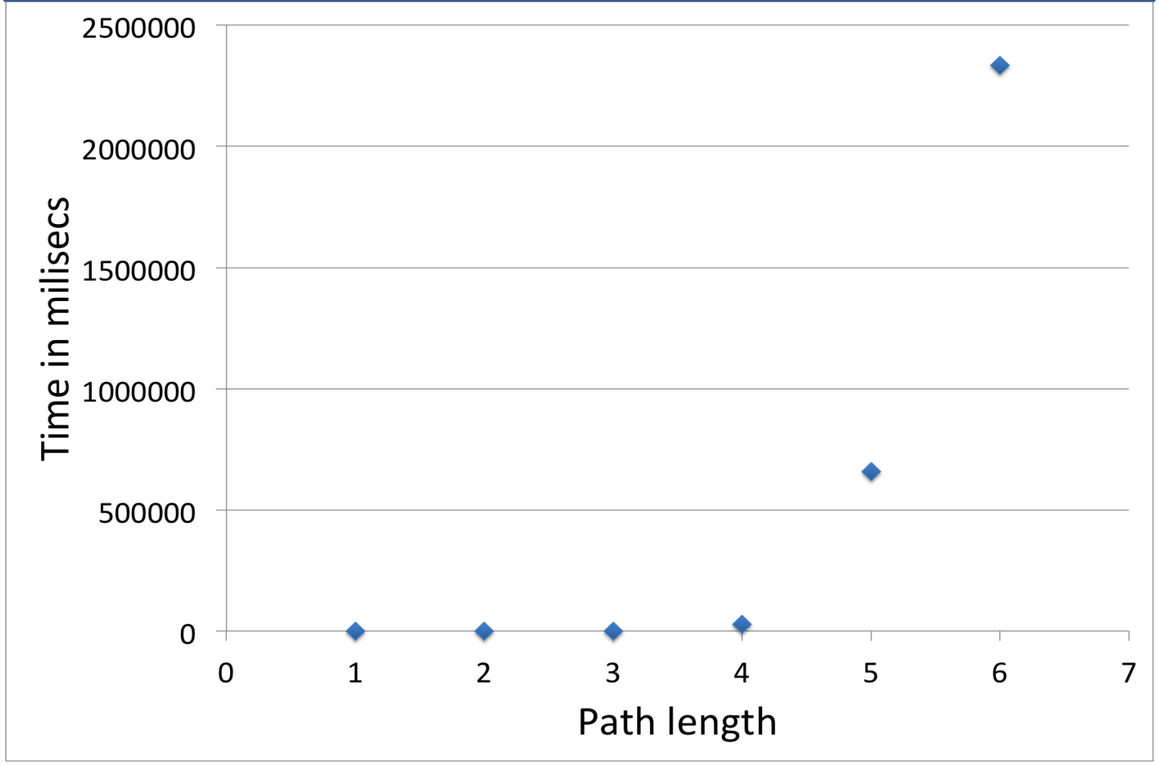
Semantic Connectivity Score – SCS (2)

- Optimization factors for Katz:
  - Exponentially many paths, measuring entity pair relatedness
  - Small world assumptions
  - Tradeoff of path length and connectivity contribution (τ=4)

#Paths with increasing length



Computation time for increasing path length



Co-occurrence-based Measure (CBM)

- Approximate number of Web resources mentioning entity pairs
- Similar to *Pointwise Mutual Information* and *Normalised Google Distance*
- Query search engines: e.g. "***Carmelo Anthony***" + "***Charlotte Bobcats***"
- Extract occurrences of each entity, and as well the entity pairs

$$CBM(e_1, e_2) = \begin{cases} 0, \text{if } count(e_1) = 0 \vee count(e_2) = 0 \\ 1, \text{if } count(e_1) = count(e_2) = count(e_1, e_2) = 1 \\ \dfrac{\log(count(e_1, e_2))}{\log(count(e_1))} \cdot \dfrac{\log(count(e_1, e_2))}{\log(count(e_2))}, \text{otherwise} \end{cases}$$

A combined entity-linking approach

- **SCS** as an exhaustive entity-linking procedure
- **CBM** –search engines to measure relatedness based on entity co-occurrence
- Complementary entity-linking results
- A combined measure, scalable and with broader coverage:

$$\alpha_{CBM+SCS}(e_i, e_j) = \begin{cases} CBM(e_i, e_j), \text{ if } CBM(e_i, e_j) > 0 \\ SCS(e_i, e_j), \text{otherwise} \end{cases}$$

Evaluation Setup

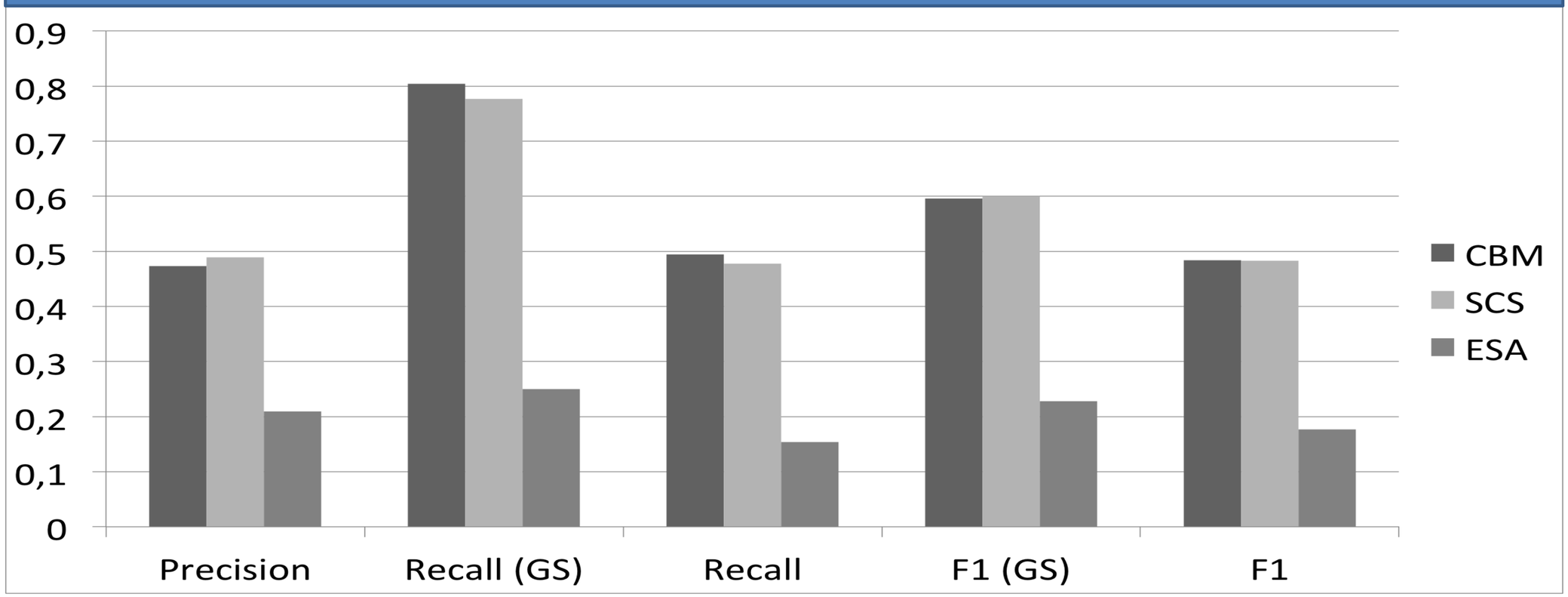
- Dataset: USAToday news
  - 40, 000 document and 80, 000 entity pairs
- *Gold standard* generated using human evaluators
  - 600 document and 1000 entity assessed pairs
- Quantify connectivity with 5-point Likert scale:
  - **correctness**: *strongly disagree* to *strongly agree*
  - **expectedness**: *extremely unexpected* to *extremely expected*
- Compare **CBM, SCS**, **ESA** entity-linking approaches
- Standard performance metrics: precision/recall/F1 measure

Entity-Linking Results

- 5-point Likert scale, entity connectivity based on gold standard:

Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
63	178	127	227	217

Precision/Recall/F1 compared against the **gold standard** for the competing approaches.



Entity-linking Results (1)

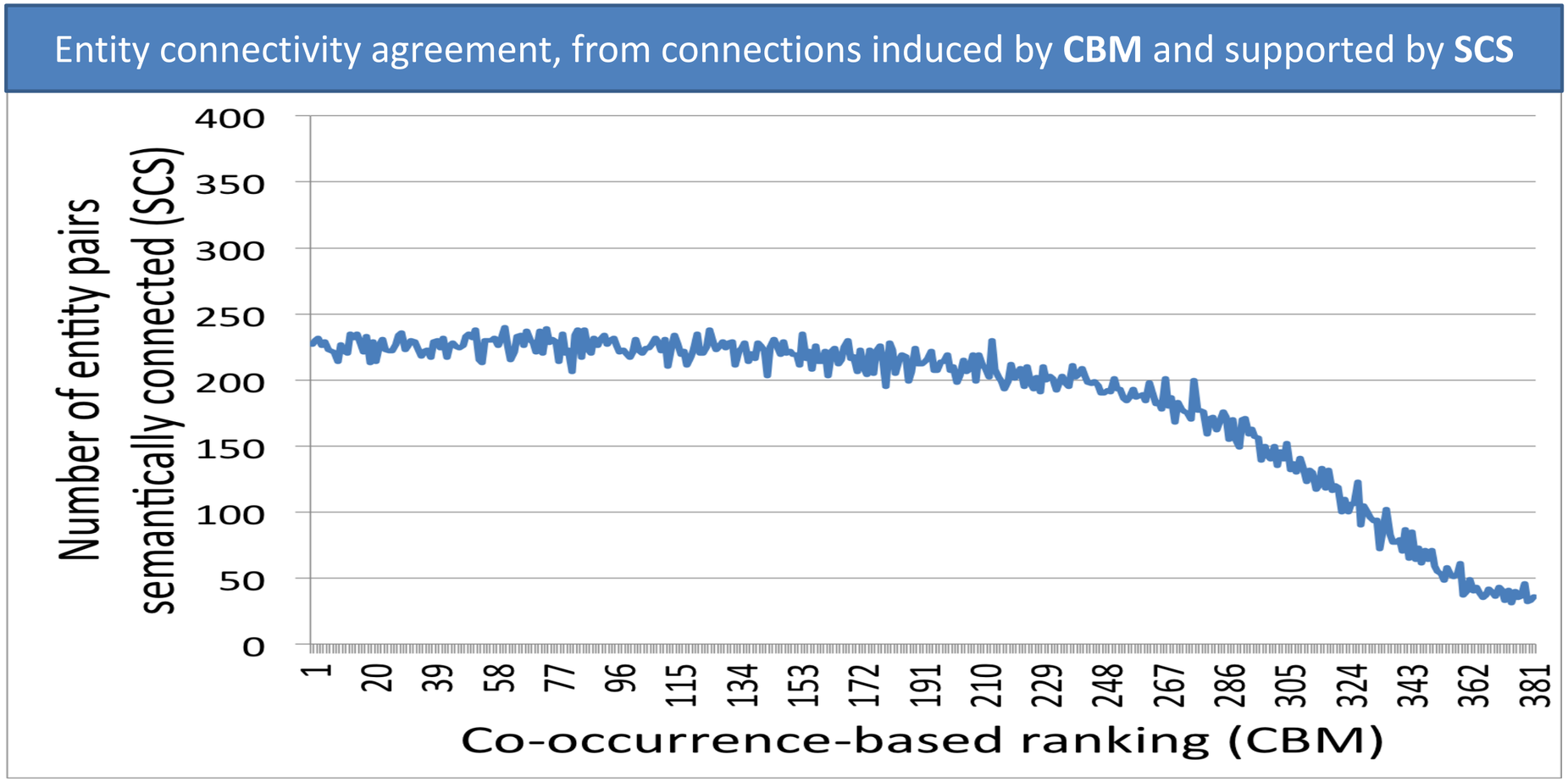
- Analysis of uncovered entity connections from competing approaches

	CBM (not in SCS)	CBM (not in ESA)	SCS (not in CBM)	SCS (not in ESA)	ESA (not in CBM)	ESA (not in SCS)
Strongly Agree	**9.5%**	**76%**	**3.1%**	**71%**	**7.9%**	**9.5%**
Agree	**12.3%**	**63.4%**	**11.2%**	**60.1%**	**8.9%**	**6.7%**
Undecided	9.4%	60.6%	6.3%	59.8%	5.5%	7.9%
Disagree	15.0%	63.0%	7.1%	53.3%	7.1%	5.3%
Strongly Disagree	18.4%	63.1%	51.6%	4.6%	4.6%	6.9%

- Expectedness of uncovered entity connections:
  - **SCS** – 25% unexpected novel entity links
  - **CBM** – 16% unexpected novel entity links

Entity-Linking Result Analysis

- Connectivity agreement: **SCS** vs. **CBM**



- Measured agreement based on Kendall's correlation coefficient:

τ	k@2	k@5	k@10
USAToday	0.40	0.47	0.52

Entity-Linking Results Analysis (1)

- Complementary entity connections between **SCS** and **CBM**

	CBM	SCS	ESA	CBM+SCS
Precision	0.32	0.34	0.16	0.34
Recall (GS)	0.81	0.78	0.23	0.90
Recall	0.52	0.51	0.15	0.58
F1 (GS)	0.46	0.47	0.19	0.50
F1	0.40	0.41	0.15	0.43

- Many entity connections labelled as "*undecided*", correct
- Examples: "Baracak Obama" and "Olympia Snowe"
  - Human evaluators, marked as not connected
  - **SCS** uncovered a connection of length 2 and more

Conclusions

- An *entity-linking* approach across disparate datasets
- Knowledge graphs, adapted and utilized to uncover entity connections via *SCS* and *CBM*
- Balanced tradeoffs between information gain and processing time for SCS
- Entity-linking *gold standard* measuring *correctness* of connectivity and *expectedness*
- Combination of SCS and CBM as scalable entity-linking approach
- Increased precision and recall based on *SCS+CBM*
- Correctly uncovered connections marked as irrelevant by human evaluators

Future Work

- Exploit semantics of edges connecting entities
- Detailed distinction of edges based on entity types
- Gold standard improvement, by showing the trace of intermediary *entities* helping uncover a connection between an entity pair
- Filtering of nodes from a knowledge graph to improve scalability

Thank you!

Questions?