DoSeR - A Knowledge-Base-Agnostic Framework for Entity Disambiguation Using Semantic Embeddings

University of Passau

Stefan Zwicklbauer

Christin Seifert

Michael Granitzer

01 June 2016

Entity Disambiguation

Depth-first Search (Research Article)

Depth-first search is an algorithm for traversing or searching tree or graph data structures. One starts at the root (selecting some arbitrary node as the root in the case of a graph) and explores as far as possible along each branch before backtracking.

The time and space analysis of DFS differs according to its application area. In theoretical computer science, DFS is typically used to traverse an entire graph, and takes time Θ(|V| + |E|) linear in the size of the graph.

Structure	Space Complexity (worst)
Stack	O(n)
List	O(n)
Tree	O(n)
Array	O(n)
Queue	O(n)

Sense

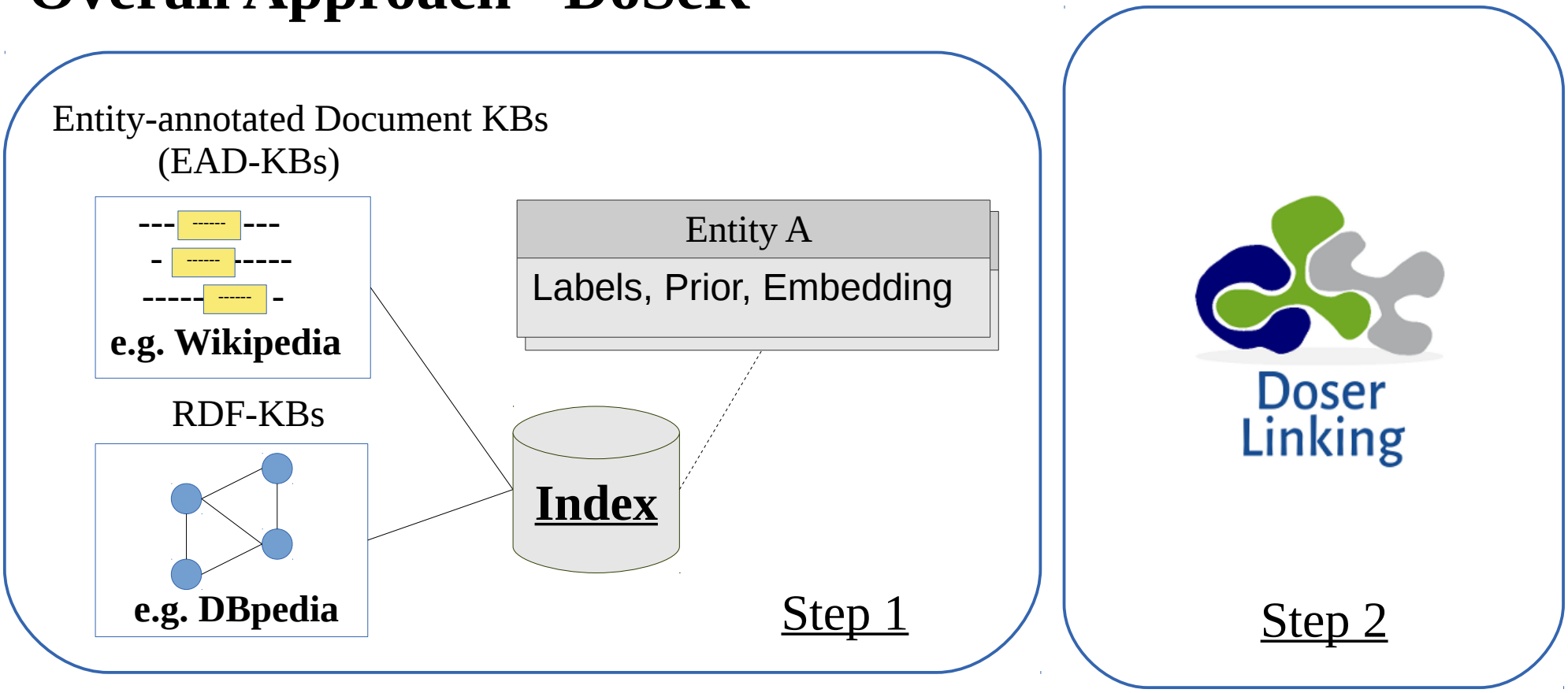
- Tree
- Tree (graph theorie)
- **Tree (data structure)**
- Tree (set theorie)
- Phylogenetic Tree
- …

The goal of Entity Disambiguation is to establish links between Surface Forms (SF) and entities in a Knowledge Base (KB) by identifying the correct semantic meaning.

Main Goal

A simple yet effective, graph-based and knowledge-base-agnostic entity disambiguation approach

Overall Approach - DoSeR



Index Creation

Entity describing information:

- Entity Label(s)
- Prior Probability
- Semantic Entity Embedding
  - Usage of Word2Vec to create Entity Embeddings
  - Accepts textual input corpora and trains its vectors according to the words ordering
  - Entities instead of words

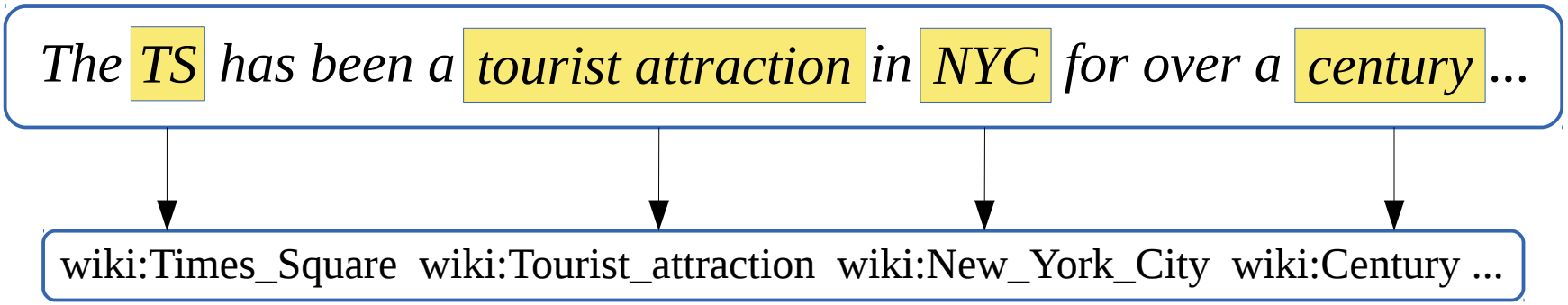
Algorithm to create W2V input corpus from EAD-KBs

For all Documents do:

- Replace all entity hyperlinks with respective entity identifier
- Replace all non-entity words and punctuations

=> Collocation of entities still maintained!

Wikipedia Example:

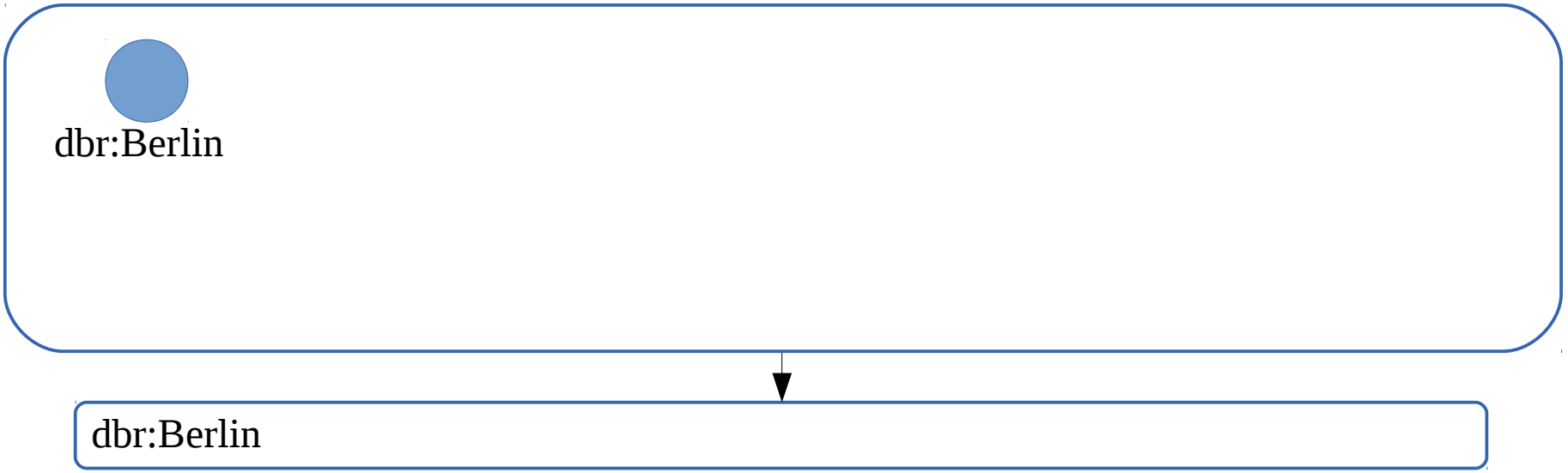


Algorithm to create W2V input corpus from RDF-KBs

Regarding RDF-KB as an undirected graph:

- Random walk from entity to entity via relations
- Normalized Inverse Edge Frequency as jump probability to any entity in the graph

DBpedia Example:

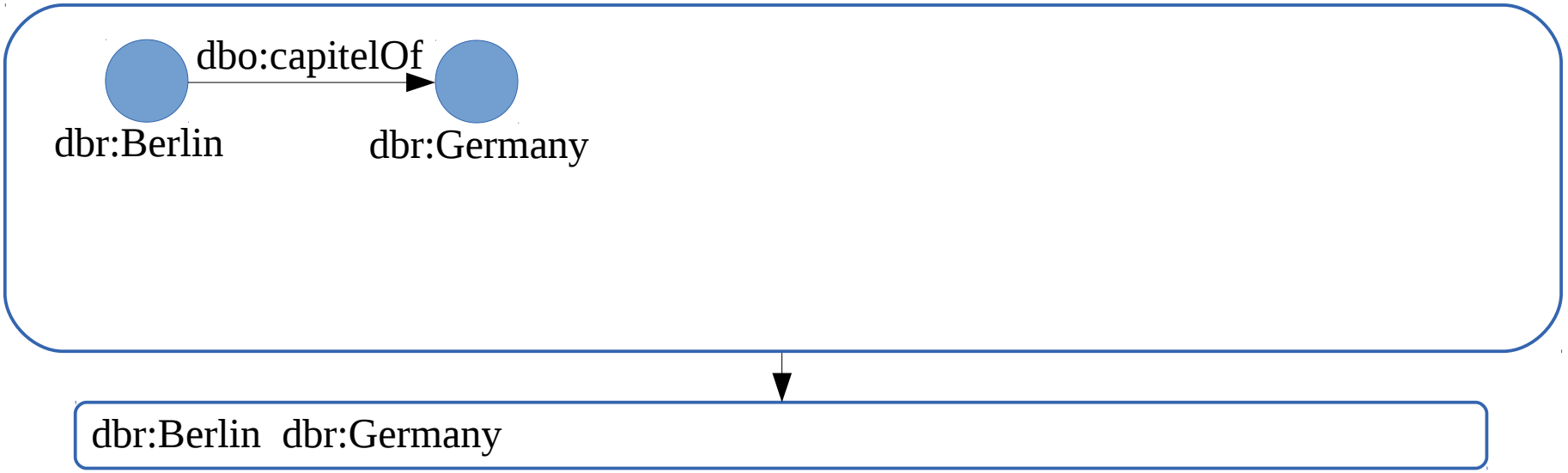


Algorithm to create W2V input corpus from RDF-KBs

Regarding RDF-KB as an undirected graph:

- Random walk from entity to entity via relations
- Normalized Inverse Edge Frequency as jump probability to any entity in the graph

DBpedia Example:

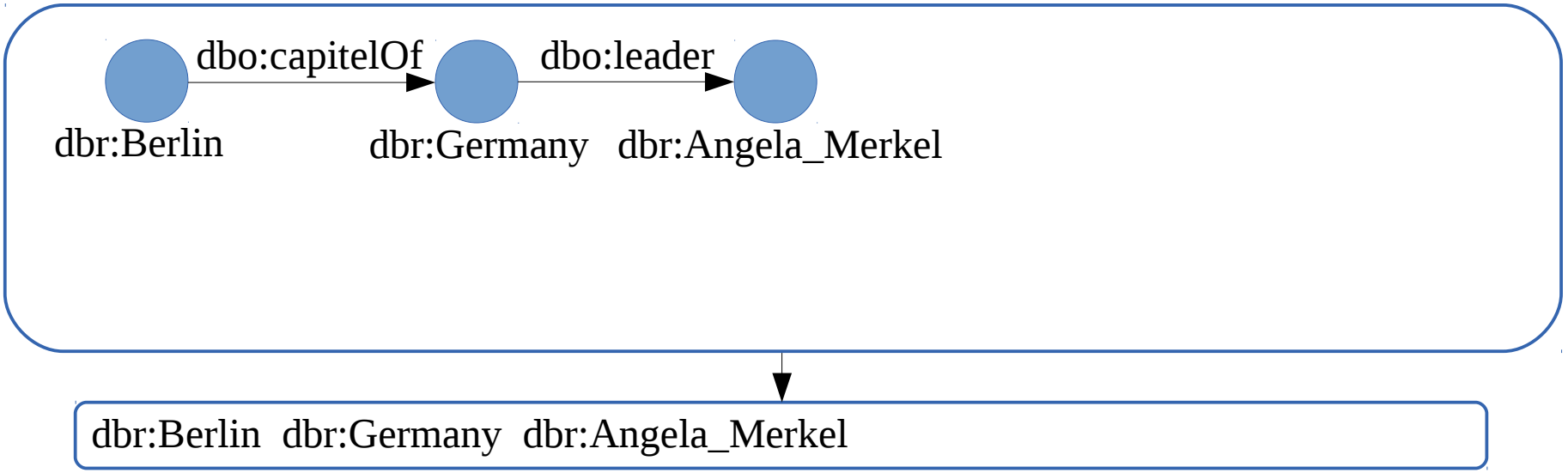


Algorithm to create W2V input corpus from RDF-KBs

Regarding RDF-KB as an undirected graph:

- Random walk from entity to entity via relations
- Normalized Inverse Edge Frequency as jump probability to any entity in the graph

DBpedia Example:

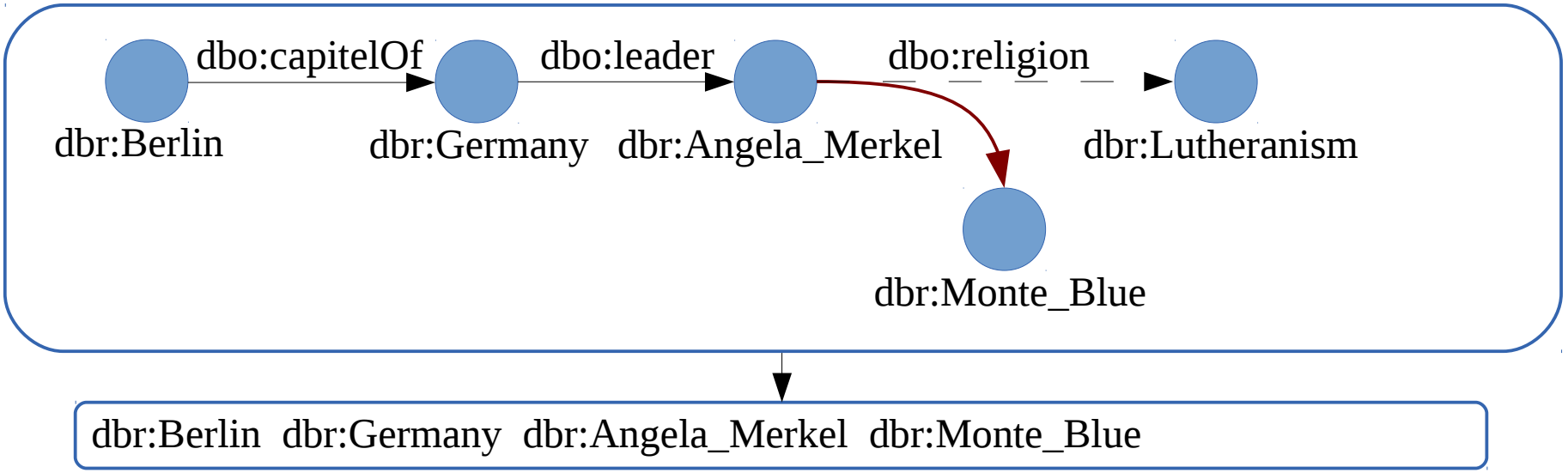


Algorithm to create W2V input corpus from RDF-KBs

Regarding RDF-KB as an undirected graph:

- Random walk from entity to entity via relations
- Normalized Inverse Edge Frequency as jump probability to any entity in the graph

DBpedia Example:

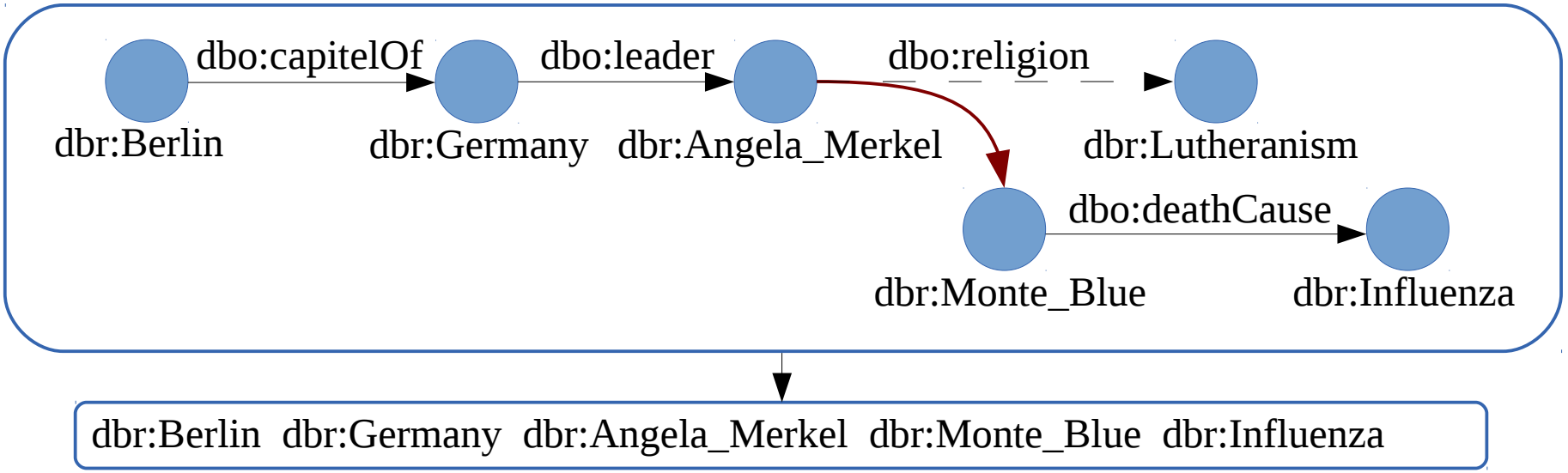


Algorithm to create W2V input corpus from RDF-KBs

Regarding RDF-KB as an undirected graph:

- Random walk from entity to entity via relations
- Normalized Inverse Edge Frequency as jump probability to any entity in the graph

DBpedia Example:



DoSeR Entity Disambiguation Algorithm

Input

- Document, SFs, Index

Example:

The TS has been a New York attraction for over a century.

DoSeR Entity Disambiguation Algorithm

Input

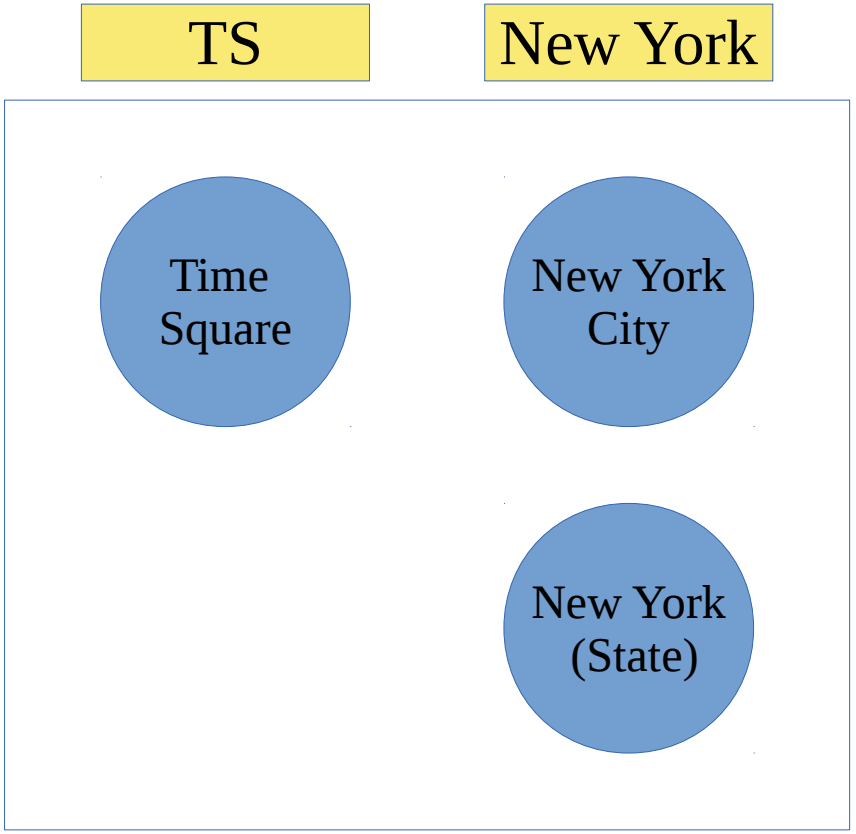
- Document, SFs, Index

Process

1. Candidates: Exact Matching, Trigram Similarity of AGDISTIS [2]

Example:

The TS has been a New York attraction for over a century.



DoSeR Entity Disambiguation Algorithm

Input

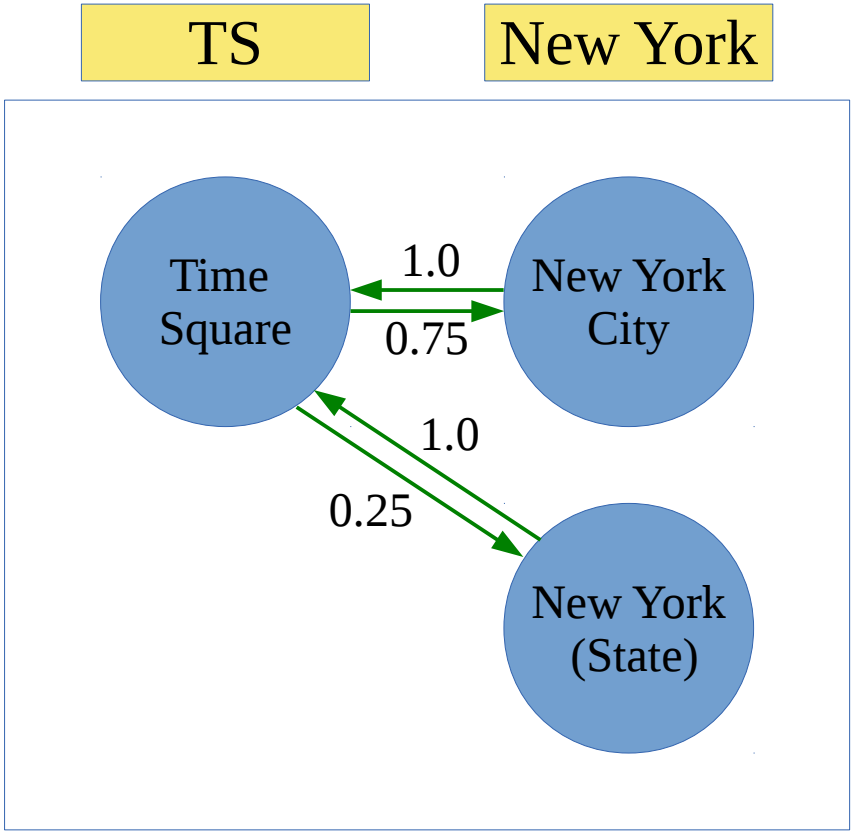
- Document, SFs, Index

Process

1. Candidates: Exact Matching, Trigram Similarity of AGDISTIS [2]
2. Create complete, directed K-partite graph without SF interconnection (Weights are normalized SR values)

Example:

The TS has been a New York attraction for over a century.



DoSeR Entity Disambiguation Algorithm

Input

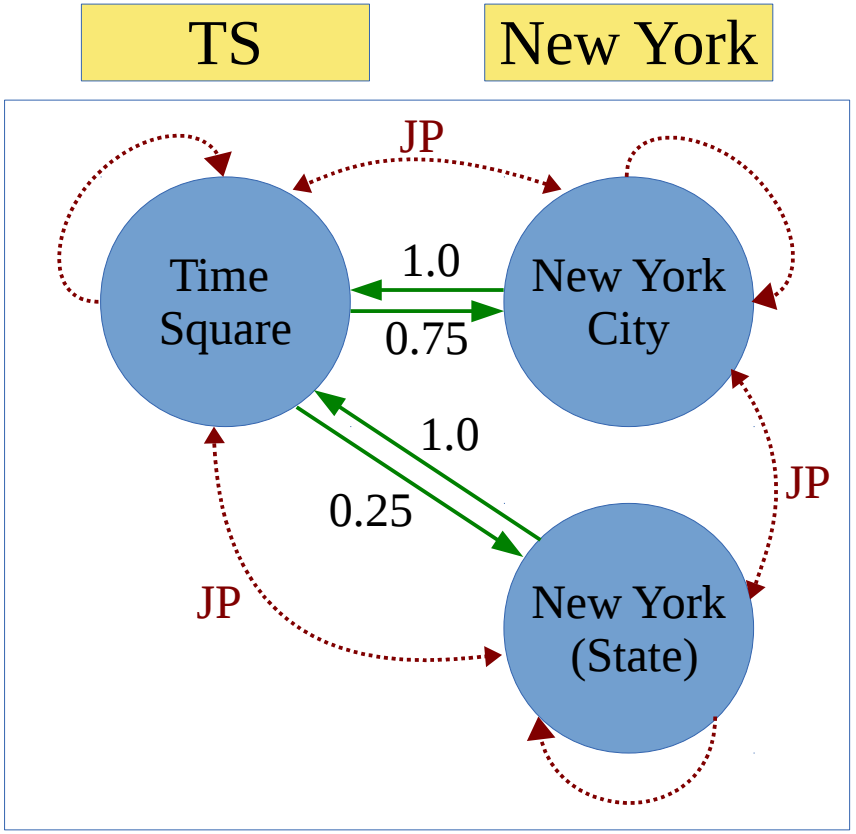
- Document, SFs, Index

Process

1. Candidates: Exact Matching, Trigram Similarity of AGDISTIS [2]
2. Create complete, directed K-partite graph without SF interconnection (Weights are normalized SR values)
3. Integrate Prior as jump probability (Perform Jump with $\alpha = 0.1$)

Example:

The TS has been a New York attraction for over a century.



DoSeR Entity Disambiguation Algorithm

Input

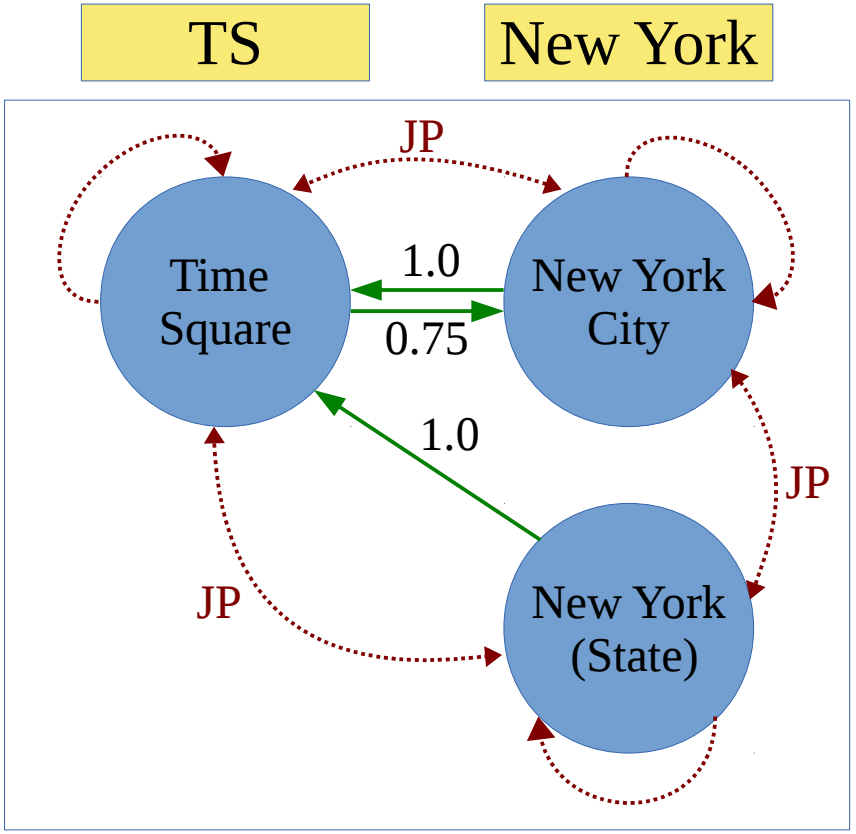
- Document, SFs, Index

Process

1. Candidates: Exact Matching, Trigram Similarity of AGDISTIS [2]
2. Create complete, directed K-partite graph without SF interconnection (Weights are normalized SR values)
3. Integrate Prior as jump probability (Perform Jump with $\alpha = 0.1$)
4. Remove 25% low probability edges

Example:

The TS has been a New York attraction for over a century.



DoSeR Entity Disambiguation Algorithm

Input

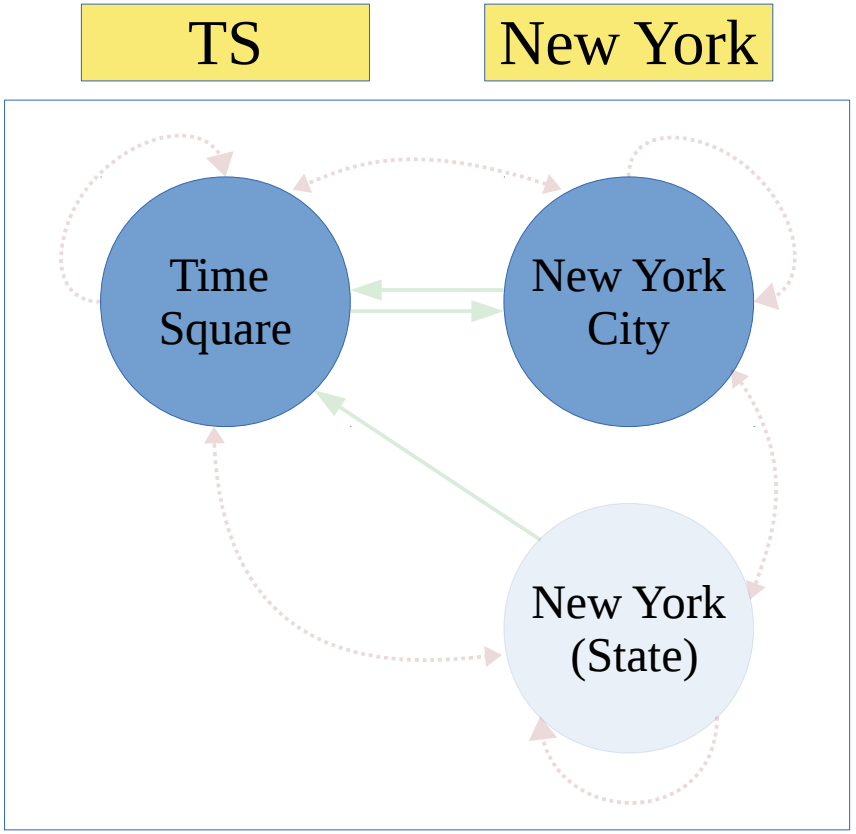
- Document, SFs, Index

Process

1. Candidates: Exact Matching, Trigram Similarity of AGDISTIS [2]
2. Create complete, directed K-partite graph without SF interconnection (Weights are normalized SR values)
3. Integrate Prior as jump probability (Perform Jump with $\alpha = 0.1$)
4. Remove 25% low probability edges

Example:

The TS has been a New York attraction for over a century.



Disambiguated Entities: Nodes with highest PageRank

DoSeR Evaluation I

- Evaluation on GERBIL v1.0 - an easy-to-use platform to evaluate disambiguation algorithms [3]
- Experiment 1:
  - Comparing DoSeR to AGDISTIS
  - Disambiguating **named entities** only
- Experiment 2:
  - Comparison against other available systems
  - Additionally leveraging Wikipedia knowledge

DoSeR Evaluation II

Experiment 1:

- DoSeR outperforms AGDISTIS on 6 out of 7 data set by about 4-5 Micro-F1 percentage points
- Semantics of entity relations can be captured by Word2Vec Embeddings
- Assuming to be robust against noisy KB information

	ACE 2004	Aida Test-B	Aquaint	MSNBC	N3-Reuters	IITB	Micro-posts (T)
Doser (DBpedia)	**0.702**	**0.616**	**0.646**	0.725	**0.731**	**0.515**	**0.489**
Doser (YAGO3)	0.679	0.608	0.611	0.735	0.725	0.454	0.454
AGDISTIS	0.658	0.582	0.596	**0.751**	0.658	0.412	0.428

Table: Micro-F1 values on 7 different data sets

DoSeR Evaluation III (+ Wikipedia)

Experiment 2:

- Training on DBpedia and Wikipedia leads to an increase of 15 percentage points Micro-F1 on average
- DoSeR (significantly) outperforms other State-of-The-Art approaches on many data sets

	ACE 2004	Aida Test-B	Aquaint	MSNBC	N3-Reuters	IITB	Micro-posts
Doser	0.681	0.597	0.638	0.719	0.700	0.497	0.469
Doser (+Wiki)	**0.864**	0.722	0.820	**0.881**	**0.727**	0.713	**0.639**
Wikifier [4]	0.824	0.776	**0.862**	0.851	0.694	**0.755**	0.586
AIDA [5]	0.741	0.806	0.534	0.796	0.571	0.277	0.412
Spotlight [6]	0.713	0.593	0.713	0.511	0.577	0.447	0.623
WAT [7]	0.800	**0.843**	0.768	0.777	0.644	0.611	0.595

Table: Micro-F1 values on 7 different data sets

Conclusion

- Presented DoSeR, a new State-of-The-Art, collective (named) entity disambiguation system
- KB-agnostic in terms of EAD-KBs and RDF-KBs
- Simple graph-based approach based on Semantic Entity Embeddings
- Presented how Semantic Entity Embeddings are generated on different KBs (e.g. DBpedia, Wikipedia)
- Future Work: Additional integration of textual context

- Code: ***https://github.com/quhfus/doser***

Thank You!

Questions?

DoSeR

https://github.com/quhfus/DoSeR

References

[1] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013)

[2] Usbeck, Ricardo, et al. "AGDISTIS-graph-based disambiguation of named entities using linked data." The Semantic Web–ISWC 2014. Springer International Publishing, 2014. 457-471.

[3] Usbeck, Ricardo, et al. "GERBIL: General Entity Annotator Benchmarking Framework." Proceedings of the 24th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2015

[4] Ratinov, Lev, et al. "Local and global algorithms for disambiguation to wikipedia." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011.

[5] Hoffart, Johannes, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum: “Robust Disambiguation of Named Entities in Text”. Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP ’11. Edinburgh, United Kingdom: ACL, 2011: pp. 782–792

[6] Mendes, Pablo N., et al. "DBpedia spotlight: shedding light on the web of documents." Proceedings of the 7th International Conference on Semantic Systems. ACM, 2011

[7] Piccinno, F., Ferragina, P.: From tagme to wat: A new entity annotator. In: First Int. Workshop on Entity Recognition/Disambiguation. pp. 55–62. ERD ’14, ACM, NY, USA (2014)