PADERBORN UNIVERSITY

Using Compositional Embeddings for Fact Checking

Ana Alexandra Morim da Silva, Michael Röder and Axel-Cyrille Ngonga Ngomo

DICE

Data Science Group
Paderborn University

International Semantic Web Conference, 2021

Introduction I

Why is fact-checking needed?

- Need for reliable data.

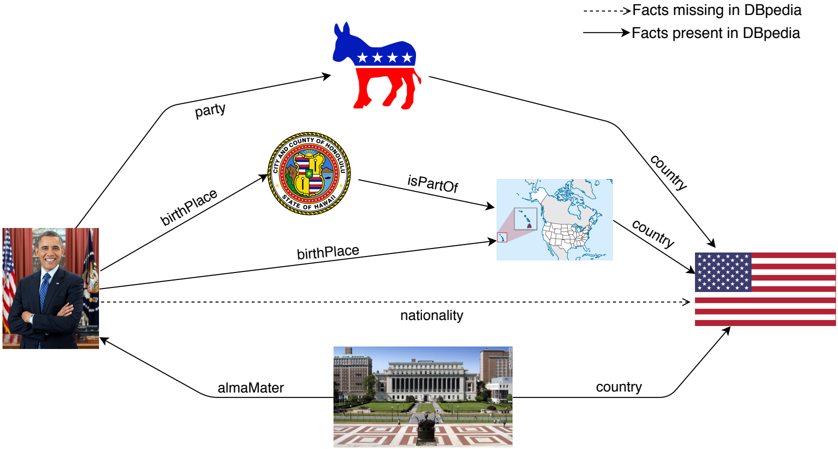


Image source: https://dice-research.org/COPAAL.

Introduction II

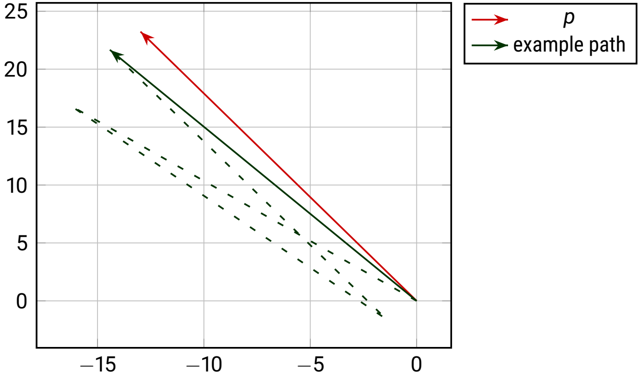
We propose **ESTHER**, a fact checker that acts on a continuous space:

- Path search leveraged by knowledge graph embedding models;
- Enables the augmentation of existing approaches.

Intuition

The embedding of the path is computed by combining the embeddings of the properties in $\vec{q}$:

$$\bigoplus_{i=1}^{|\vec{q}|} e(q_i) \approx e(p) \tag{1}$$



Search requirements

- Cost function:

$$\min(\varepsilon) = \min\left(||e(\vec{q}) - e(p)|| + \eta|\vec{q}|\right) \tag{2}$$

- Search space defined by a binary matrix $\mathcal{M}_{i,j}$:

$$\mathcal{M}_{i,j} = \begin{cases} 1, & \text{if properties } p_i \text{ and } p_j \text{ can be adjacent} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Search modes

- 5 different property-adjacency modes:

Strict equality (S) : $$\mathcal{M}_{i,j} = 1 \Leftrightarrow R'(p_i) = D'(p_j) \quad (4)$$

Search modes

- 5 different property-adjacency modes:

Strict equality (S) : $$\mathcal{M}_{i,j} = 1 \Leftrightarrow R'(p_i) = D'(p_j) \tag{4}$$

Subsumed (SU) : $$\mathcal{M}_{i,j} = 1 \Leftrightarrow R(p_i) \supseteq D(p_j) \tag{5}$$

Search modes

- 5 different property-adjacency modes:

$$\text{Strict equality (S)}: \quad \mathcal{M}_{i,j} = 1 \Leftrightarrow R'(p_i) = D'(p_j) \tag{4}$$

$$\text{Subsumed (SU)}: \quad \mathcal{M}_{i,j} = 1 \Leftrightarrow R(p_i) \supseteq D(p_j) \tag{5}$$

$$\text{Non-disjoint (ND)}: \quad \mathcal{M}_{i,j} = 1 \Leftrightarrow R'(p_i) \cap D'(p_j) \neq \emptyset \tag{6}$$

Search modes

- 5 different property-adjacency modes:

Strict equality (S) : $$\mathcal{M}_{i,j} = 1 \Leftrightarrow R'(p_i) = D'(p_j) \quad (4)$$

Subsumed (SU) : $$\mathcal{M}_{i,j} = 1 \Leftrightarrow R(p_i) \supseteq D(p_j) \quad (5)$$

Non-disjoint (ND) : $$\mathcal{M}_{i,j} = 1 \Leftrightarrow R'(p_i) \cap D'(p_j) \neq \emptyset \quad (6)$$

Non-disjoint subsumption (NDS) : $$\mathcal{M}_{i,j} = 1 \Leftrightarrow R(p_i) \cap D(p_j) \neq \emptyset \quad (7)$$

PADERBORN UNIVERSITY

Search modes

DICE

- 5 different property-adjacency modes:

$$\text{Strict equality (S)}: \quad \mathcal{M}_{i,j} = 1 \Leftrightarrow R'(p_i) = D'(p_j) \tag{4}$$

$$\text{Subsumed (SU)}: \quad \mathcal{M}_{i,j} = 1 \Leftrightarrow R(p_i) \supseteq D(p_j) \tag{5}$$

$$\text{Non-disjoint (ND)}: \quad \mathcal{M}_{i,j} = 1 \Leftrightarrow R'(p_i) \cap D'(p_j) \neq \emptyset \tag{6}$$

$$\text{Non-disjoint subsumption (NDS)}: \quad \mathcal{M}_{i,j} = 1 \Leftrightarrow R(p_i) \cap D(p_j) \neq \emptyset \tag{7}$$

$$\text{Irrelevant (I)}: \quad \forall p_i, p_j \in \mathbb{P}^*_{\mathcal{G}}, \mathcal{M}_{i,j} = 1 \tag{8}$$

Path search

1. Initialize priority queue;
   - $D(p) = D(p_i)$
2. Iterate queue in a sorted manner.
   - Stop properties:

$$R(p) = R(p_{i-1}) \tag{9}$$

   - Intermediate properties:

$$D(p_i) = R(p_{i-1}) \tag{10}$$

Subject to

- maximum path length k;
- maximum number of paths N;
- cycle-free paths indicator.

Statement validation at runtime

1. Corroborative path check between subject and object;
2. Summarize the found paths' scores.

$$\tau = \sqrt[3]{\frac{1}{|Z|} \sum_{\zeta_{p,\vec{q}} \in Z} \zeta_{p,\vec{q}}^{3}} \tag{11}$$

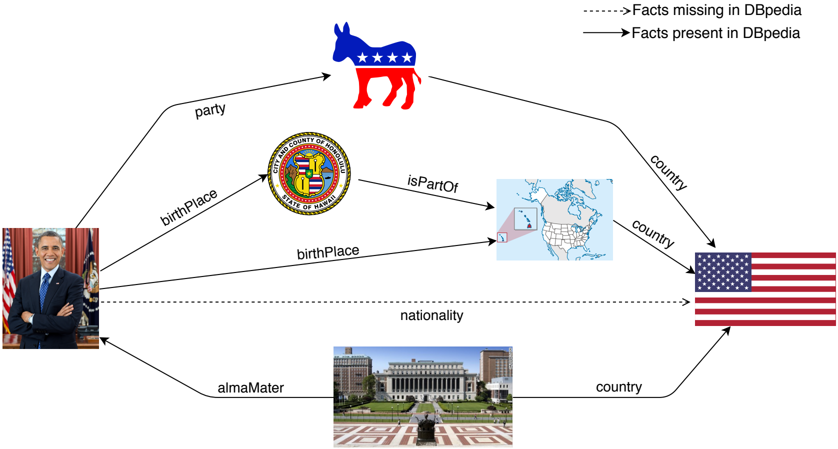


Image source: https://dice-research.org/COPAAL.

ESTHER

Path found	A^* cost	PNPMI
$\xrightarrow{\texttt{place\_of\_birth}} \xleftarrow{\texttt{marriage.location\_of\_ceremony}} \xrightarrow{\texttt{nationality}}$	7.87	0.47
$\xrightarrow{\texttt{people.place\_lived.location}} \xleftarrow{\texttt{place\_of\_birth}} \xrightarrow{\texttt{sibling}} \xrightarrow{\texttt{nationality}}$	9.30	0.27
$\xrightarrow{\texttt{languages}} \xrightarrow{\texttt{countries\_spoken\_in}}$	10.88	0.07

- Example paths for the predicate `nationality` found by ESTHER in the FB15k-237 dataset;
- The veracity score is higher for lower search scores.

Experiment setup

Datasets:

- FB15K-237 [1];
- WN18RR [2].

KGE models:

- TransE [3];
- RotatE [4];
- DensE [5].

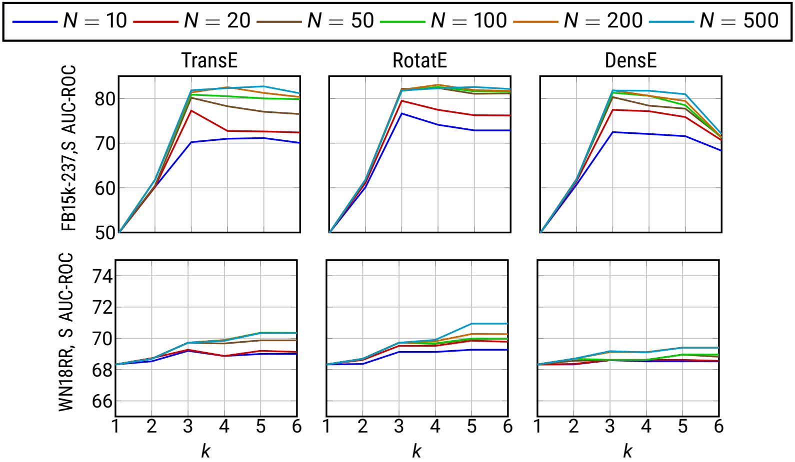
Evaluation

Different configurations experiment I

	KGE	Mode	Loops	k	N	AUC-ROC
FB15k-237	TransE	S	Y	4	200	82.53
	RotatE	S	Y	4	200	**83.07**
	DensE	S	Y	3	200	81.82
WN18RR	TransE	I	Y	3	500	**77.55**
	RotatE	I	Y	5	500	73.15
	DensE	I	Y	2	500	71.19

Evaluation

Different configurations experiment II



Evaluation

Comparison with other approaches

	AUC-ROC	
Approach	FB15k-237	WN18RR
ESTHER	83.07	77.55
COPAAL [6]	77.42	68.11
KS [7]	**87.59**	**86.44**
Katz [8]	82.80	69.96
Pathent [9]	73.46	79.98
Simrank [10]	40.07	44.15
AdamicAdar [11]	72.12	59.86
Jaccard [12]	38.56	42.34
Degree product [13]	77.11	65.57
PredPath [14]	69.87	80.20
PRA [15]	08.53	71.80

Evaluation

Combinations with ESTHER

	AUC-ROC	
Approach	FB15k-237	WN18RR
COPAAL	87.12(+09.70)	79.38(+11.27)
KS	89.97(+02.38)	**94.92**(+08.48)
Katz	86.30(+03.50)	86.22(+16.26)
Pathent	84.75(+11.29)	86.94(+06.96)
Simrank	81.60(+41.53)	82.47(+38.32)
AdamicAdar	85.36(+13.24)	84.22(+24.36)
Jaccard	82.91(+44.35)	87.18(+44.84)
Degree product	83.28(+06.17)	87.43(+21.86)
PredPath	83.76(+13.89)	82.20(+02.00)
PRA	**97.44**(+88.91)	75.35(+03.55)

Evaluation

Runtime experiment

Approach		FB15k-237	WN18RR
ESTHER	Pre-processing	1 684	99
	Fact checking	375	1166
	Sum	2 059	1265
COPAAL		408	67
KS		9 702	3 259
Katz		11 227	722
Pathent		2 324	690
Simrank		4 003	3 219
AdamicAdar		2	1
Jaccard		2	1
Degree product		2	1
PredPath		2 400	660
PRA		10 654	13 046

Summary

- First path-based approach on continuous representations;
- ESTHER as a standalone system can be competitive;
- The performance improved when combined with other approaches.

https://github.com/dice-group/esther/

This work has been supported by the German Federal Ministry of Education and Research (BMBF) within the EuroStars project E!113314 FROCKG under the grant no 01QE19418. This work has been supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860801.

References I

[1] K. Toutanova, D. Chen, P. Pantel, H. Poon, P. Choudhury, and M. Gamon, "Representing text for joint embedding of text and knowledge bases," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1499–1509, Sept. 2015.

[2] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, "Convolutional 2d knowledge graph embeddings," *CoRR*, vol. abs/1707.01476, 2017.

[3] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in neural information processing systems*, pp. 2787–2795, 2013.

[4] Z. Sun, Z. Deng, J. Nie, and J. Tang, "Rotate: Knowledge graph embedding by relational rotation in complex space," *CoRR*, vol. abs/1902.10197, 2019.

[5] H. Lu and H. Hu, "Dense: An enhanced non-abelian group representation for knowledge graph embedding," 2020.

[6] Z. H. Syed, M. Röder, and A.-C. Ngonga Ngomo, "Unsupervised discovery of corroborative paths for fact validation," in *The Semantic Web – ISWC 2019*, pp. 630–646, 2019.

[7] P. Shiralkar, A. Flammini, F. Menczer, and G. L. Ciampaglia, "Finding streams in knowledge graphs to support fact checking," in *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 859–864, IEEE, 2017.

[8] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, 1953.

References II

[9] Z. Xu, C. Pu, and J. Yang, "Link prediction based on path entropy," *Physica A: Statistical Mechanics and its Applications*, vol. 456, pp. 294–301, 2016.

[10] G. Jeh and J. Widom, "Simrank: A measure of structural-context similarity," in *Proceedings of the Eighth ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining*, 2002.

[11] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social Networks*, vol. 25, no. 3, 2003.

[12] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *Proceedings of the Twelfth Intern. Conf. on Information and Knowledge Management*, 2003.

[13] B. Shi and T. Weninger, "Fact checking in large knowledge graphs - A discriminative predicate path mining approach," *CoRR*, vol. abs/1510.05911, 2015.

[14] B. Shi and T. Weninger, "Discriminative predicate path mining for fact checking in knowledge graphs," *Knowledge-based systems*, vol. 104, pp. 123–133, 2016.

[15] N. Lao and W. W. Cohen, "Relational retrieval using a combination of path-constrained random walks," *Machine learning*, vol. 81, no. 1, pp. 53–67, 2010.

[16] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "Pathsim: Meta path-based top-k similarity search in heterogeneous information networks," *Proceedings of the VLDB Endowment*, vol. 4, no. 11, 2011.

References III

[17] M. Zhao, T. W. Chow, Z. Zhang, and B. Li, “Automatic image annotation via compact graph based semi-supervised learning,” *Knowledge-Based Systems*, vol. 76, pp. 148–165, 2015.