A Scalable approach to compute Semantic Relatedness using Semantic Web Data

Dennis Diefenbach, Ricardo Usbeck, Kamal Singh, Pierre Maret

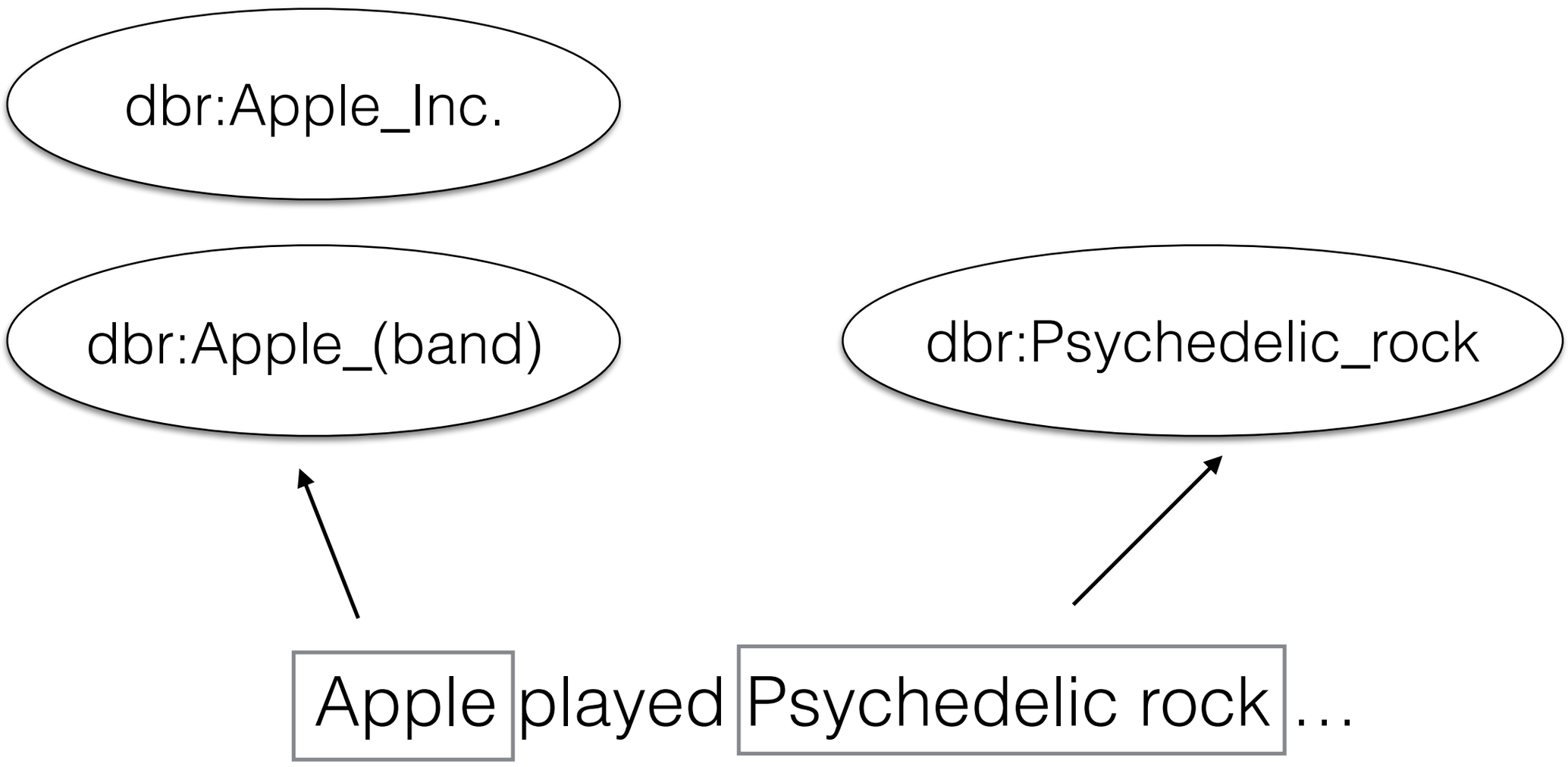
LABORATOIRE HUBERT CURIEN
UMR • CNRS • 5516 • SAINT-ETIENNE

Semantic relatedness between Named Entities

dbr:Apple_Inc. vs. dbr:Psychedelic_rock

dbr:Apple_(band) vs. dbr:Psychedelic_rock

Semantic relatedness for Entity Linking



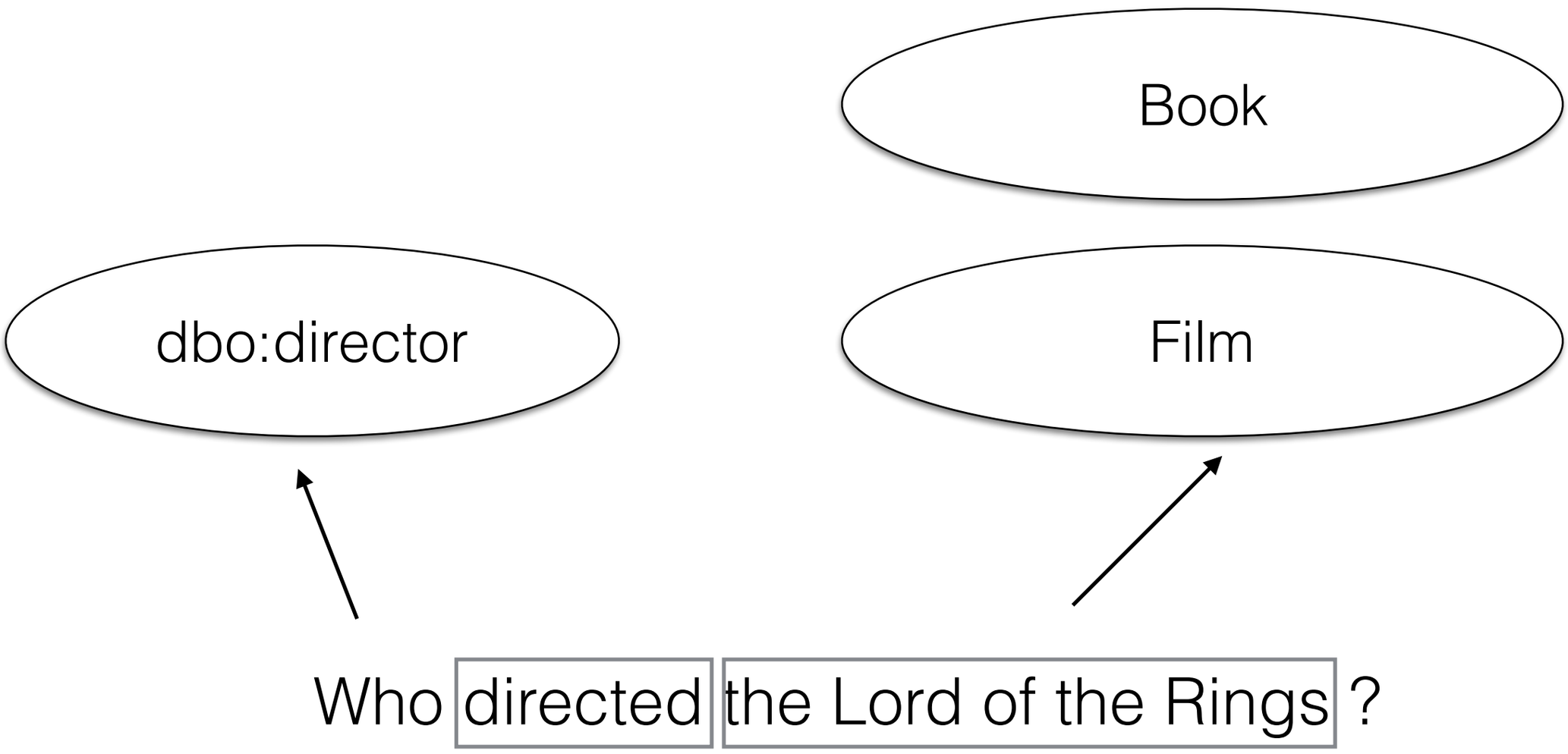
Semantic relatedness including relations …

dbr:The_Lord_of_the_Rings **vs.** dbo:director

dbr:The_Lord_of_the_Rings_(film_series) **vs.** dbo:director

dbo:child **vs.** dbo:founded

Semantic relatedness for Question Answering

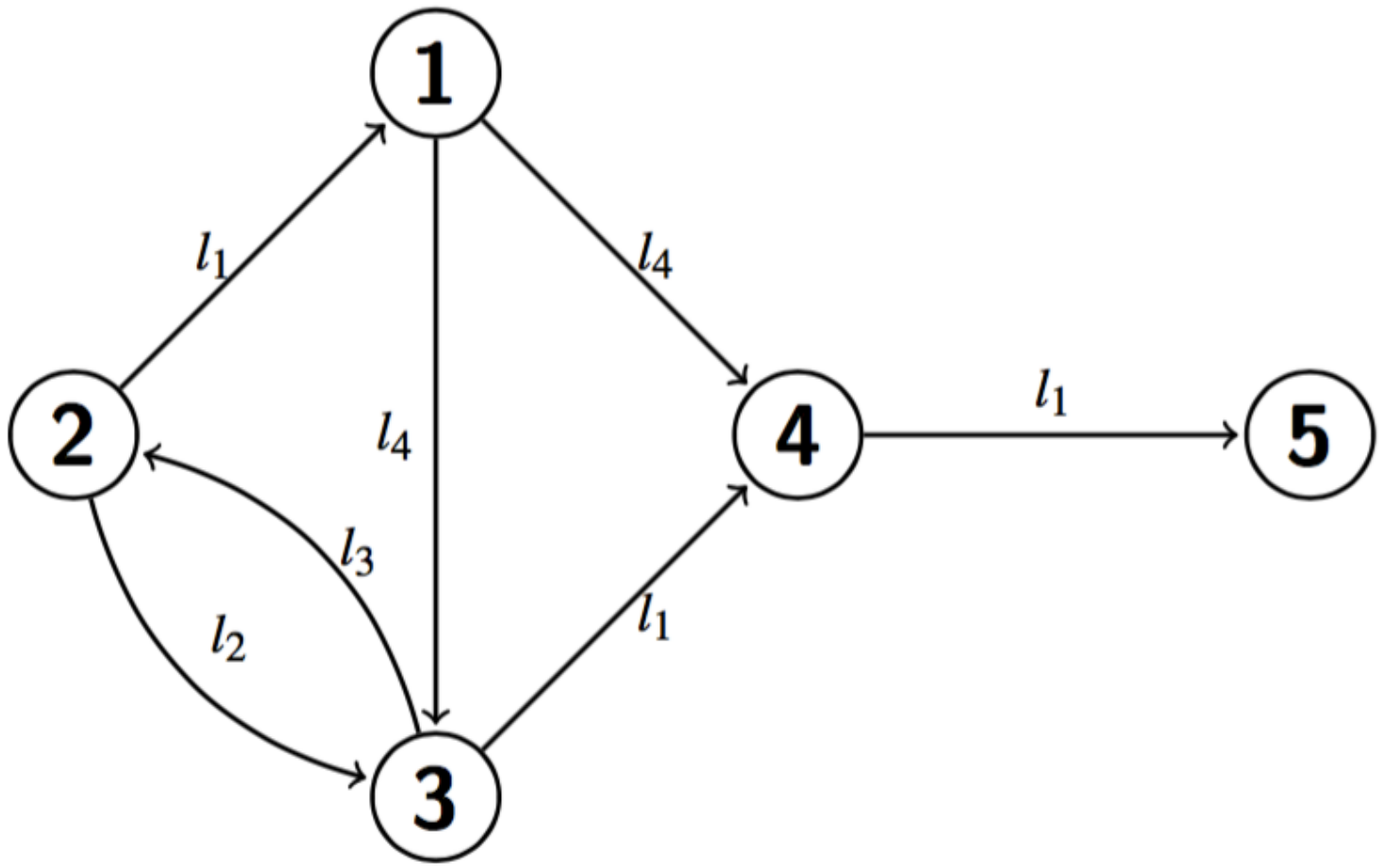


Why is scalability important for semantic relatedness?

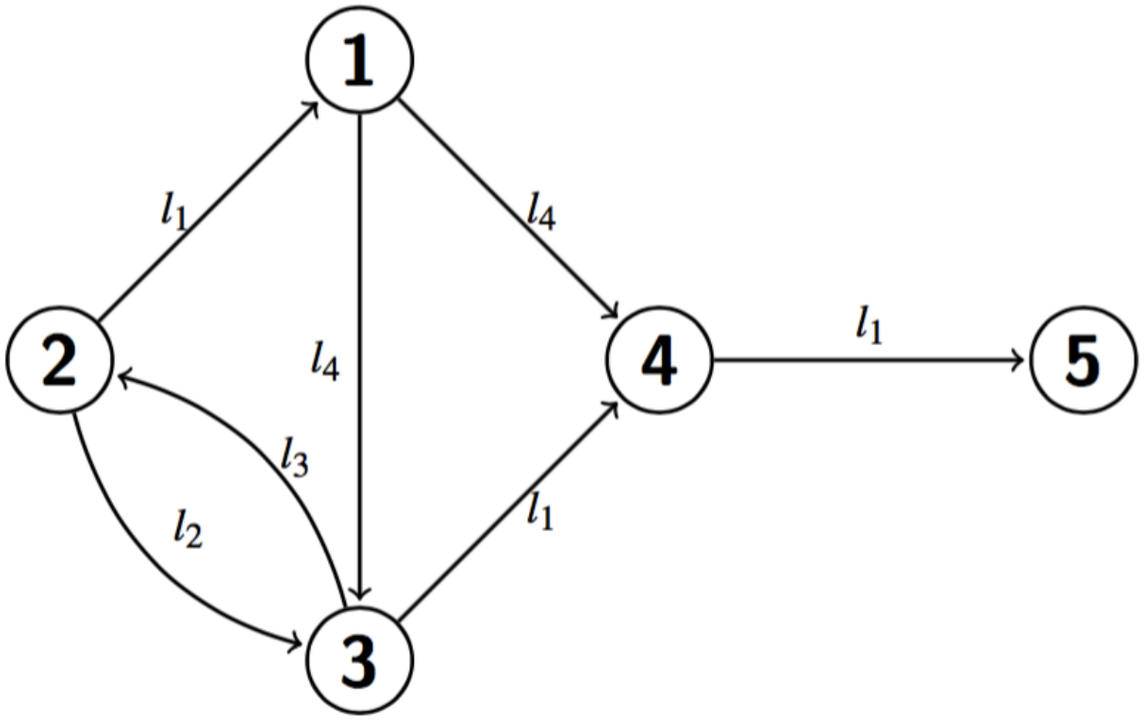
"What are the names of the children of Obama and Michelle?"

In DBpedia there are 979 instances which contain in their label "Obama" and 1204 instances which contain in their label "Michelle"

How to compute semantic relatedness using ontologies?

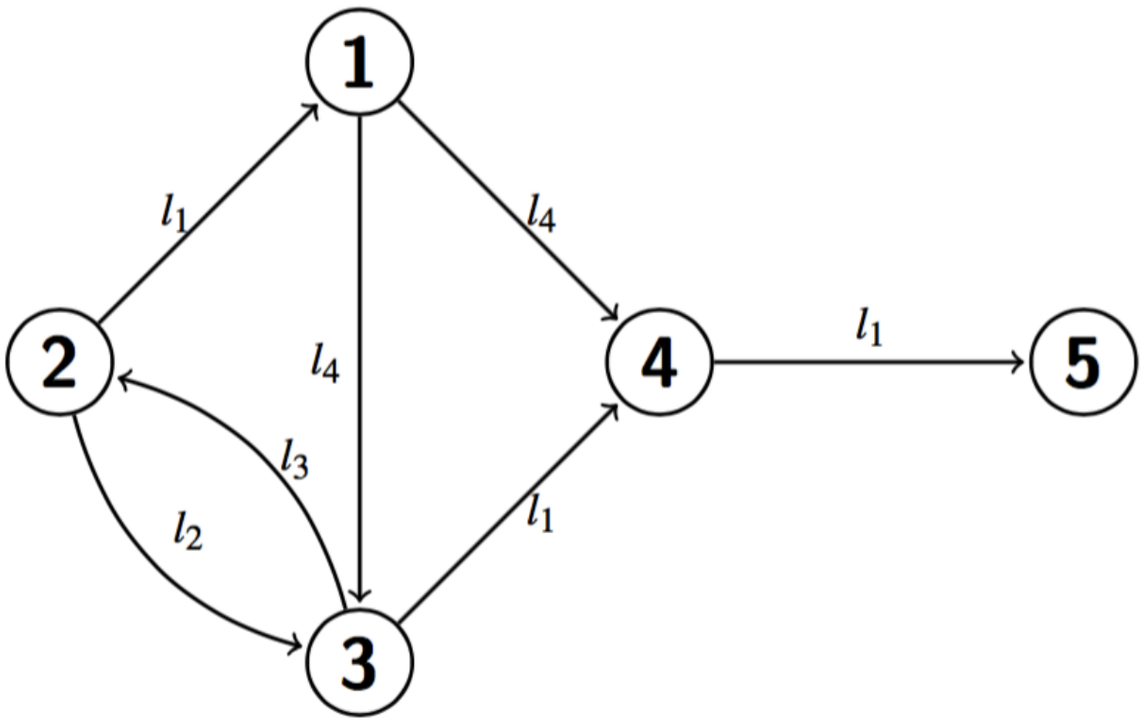


How to calculate it?



$$A_G = \begin{pmatrix} 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

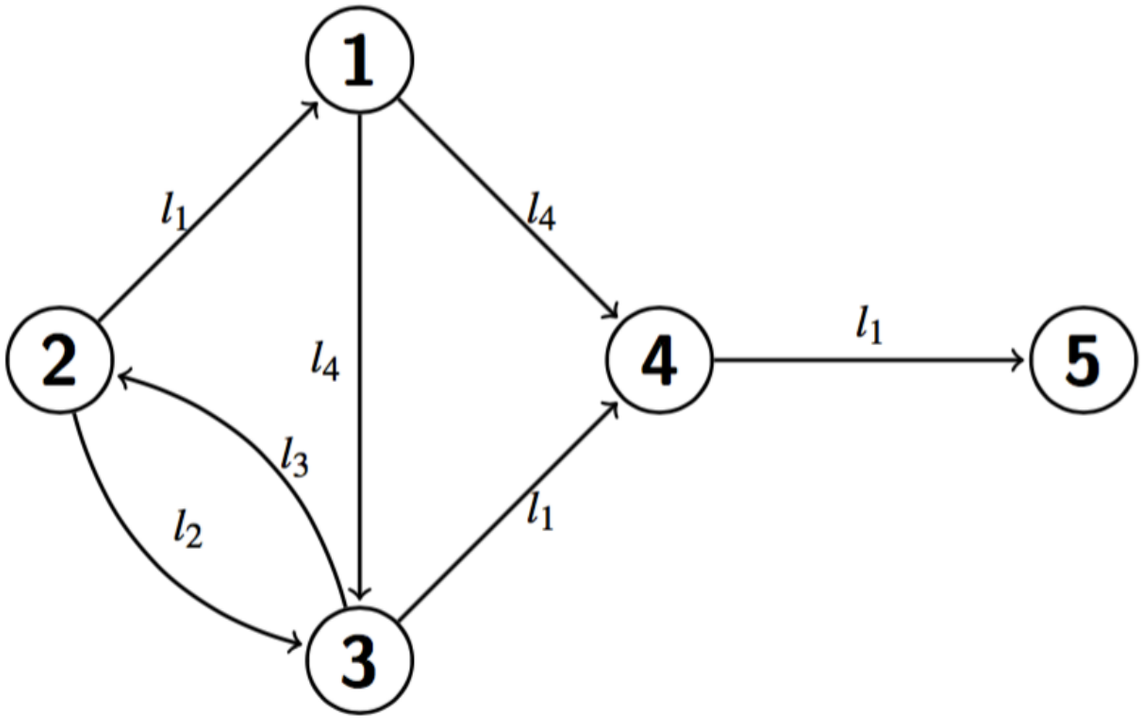
How to calculate it?



$$A_G^2 = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & \mathbf{2} & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

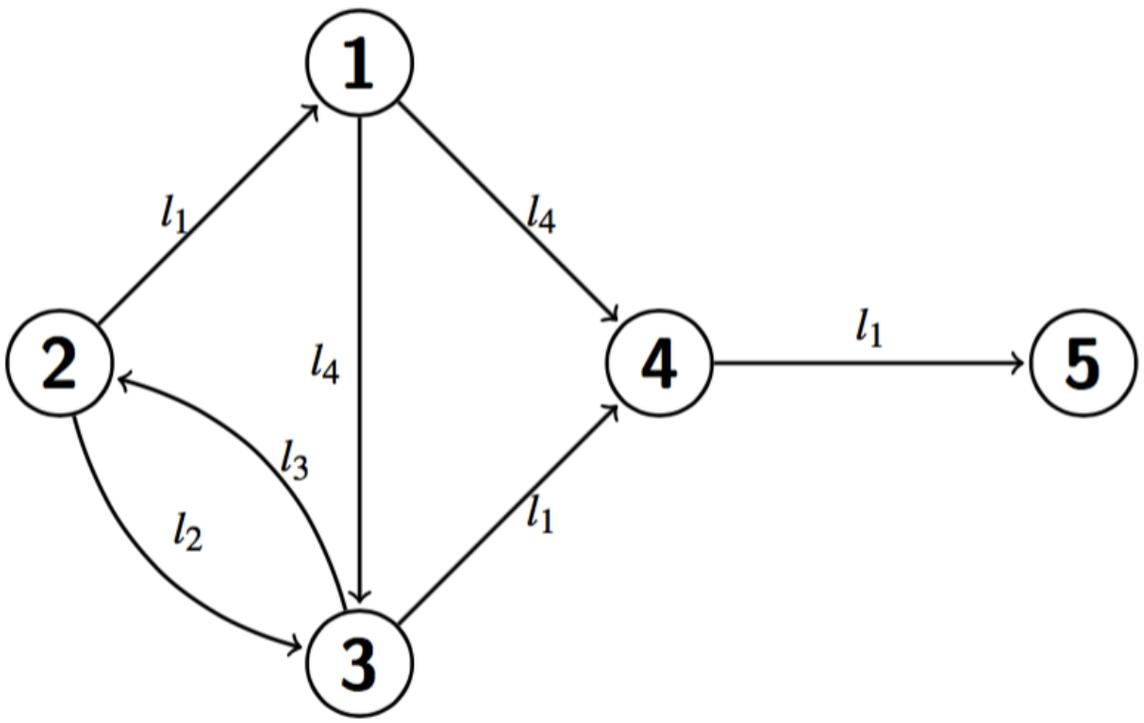
$$A_G^3 = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 2 \\ 0 & 1 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

How to calculate it?



$$D_1 = \left(\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 3 & 2 & 1 & 1 & 2 \\ 2 & 1 & 2 & 1 & 2 & 3 \\ 3 & 2 & 1 & 2 & 1 & 2 \\ 4 & 0 & 0 & 0 & 0 & 1 \\ 5 & 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

Including relations



$$D_2 = \left(\begin{array}{c|ccccc|cccc} & 1 & 2 & 3 & 4 & 5 & l_1 & l_2 & l_3 & l_4 \\ \hline 1 & 6 & 4 & 2 & 2 & 4 & 3 & 5 & 3 & 1 \\ 2 & 2 & 4 & 2 & 4 & 6 & 1 & 1 & 3 & 3 \\ 3 & 4 & 2 & 4 & 2 & 4 & 1 & 3 & 1 & 5 \\ 4 & 0 & 0 & 0 & 0 & 2 & 1 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline l_1 & 1 & 5 & 3 & 1 & 1 & 2 & 6 & 4 & 2 \\ l_2 & 5 & 3 & 1 & 3 & 5 & 2 & 4 & 2 & 6 \\ l_3 & 3 & 1 & 3 & 5 & 0 & 2 & 2 & 4 & 4 \\ l_4 & 5 & 3 & 1 & 1 & 3 & 2 & 4 & 2 & 6 \end{array}\right)$$

Including relations …

$$N_{2h+1} = \left(\begin{array}{c|c} 0 & A_G^h \cdot R_{G,L,1} \\ \hline R_{G,L,1} \cdot A_G^h & 0 \end{array}\right)$$

$$N_{2h} = \left(\begin{array}{c|c} A_G^h & 0 \\ \hline 0 & R_{G,L,2} \cdot A_G^{h-1} \cdot R_{G,L,1} \end{array}\right)$$

$$D_2 = \left(\begin{array}{c|ccccc|cccc} & 1 & 2 & 3 & 4 & 5 & l_1 & l_2 & l_3 & l_4 \\ \hline 1 & 6 & 4 & 2 & 2 & 4 & 3 & 5 & 3 & 1 \\ 2 & 2 & 4 & 2 & 4 & 6 & 1 & 1 & 3 & 3 \\ 3 & 4 & 2 & 4 & 2 & 4 & 1 & 3 & 1 & 5 \\ 4 & 0 & 0 & 0 & 0 & 2 & 1 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline l_1 & 1 & 5 & 3 & 1 & 1 & 2 & 6 & 4 & 2 \\ l_2 & 5 & 3 & 1 & 3 & 5 & 2 & 4 & 2 & 6 \\ l_3 & 3 & 1 & 3 & 5 & 0 & 2 & 2 & 4 & 4 \\ l_4 & 5 & 3 & 1 & 1 & 3 & 2 & 4 & 2 & 6 \end{array}\right)$$

The important thing: it scales

N	dump	Server	Relations	Depth	Dictionary	Matrix	Total
1	dump-it	laptop	no	3	9.7 min 1.6 GB	1.5 min 481 MB	11.2 min 2.1 GB
2	dump-it	laptop	yes	6	9.7 min 1.6 GB	6.6 min 1.9 GB	16.3 min 3.5 GB
3	dump-en	lahc-2	no	3	7.3 min 4.0 GB	5.6 min 5.3 GB	12.9 min 9.3 GB
4	dump-en	lahc-2	yes	4	9.9 min 4.0 GB	7.2 min 0.5 GB	17.1 min 4.5 GB
5	dump-en	lahc-6	yes	6	5.0 min 8.0 GB	3.6 min 20.5 GB	8.6 min 28.5 GB
5	dump-en	lahc-6	yes	6	5.0 min 8.0 GB	3.6 min 20.5 GB	8.6 min 28.5 GB
6	dump-en-P	lahc-6	no	3	11.4 min 8.0 GB	24.8 min 21.2 GB	36.2 min 29.2 GB

Modifying the EL tool AGDISTIS

Annotator	Dataset	Micro-F1	Micro-Precision	Micro-Recall	avg time ms/documents
AGDISTIS	ACE2004	0.5987	0.6020	0.5948	311.1964
AGDISTIS-mod	ACE2004	0.6209	0.6209	0.6209	238.3684
AGDISTIS	AIDA/CoNLL-Complete	0.4849	0.4851	0.4846	660.9863
AGDISTIS-mod	AIDA/CoNLL-Complete	0.4883	0.4883	0.4883	492.8392
AGDISTIS	AIDA/CoNLL-Test A	0.4884	0.4884	0.4884	736.7731
AGDISTIS-mod	AIDA/CoNLL-Test A	0.4913	0.4913	0.4913	559.7454
AGDISTIS	AIDA/CoNLL-Test B	0.4525	0.4525	0.4525	715.8788
AGDISTIS-mod	AIDA/CoNLL-Test B	0.4583	0.4583	0.4583	482.7229
AGDISTIS	AIDA/CoNLL-Training	0.492	0.4925	0.4915	625.1439
AGDISTIS-mod	AIDA/CoNLL-Training	0.4583	0.4583	0.4583	487.1205
AGDISTIS	AQUAINT	0.4542	0.4586	0.4498	611.3061
AGDISTIS-mod	AQUAINT	0.4704	0.4704	0.4704	535.8
AGDISTIS	DBpediaSpotlight	0.2485	0.2516	0.2455	194.807
AGDISTIS-mod	DBpediaSpotlight	0.2606	0.2606	0.2606	188.4655
AGDISTIS	IITB	0.4648	0.4665	0.4632	4,147.2621
AGDISTIS-mod	IITB	0.4664	0.4664	0.4664	3,602.75
AGDISTIS	KORE50	0.3228	0.3262	0.3194	391.5714
AGDISTIS-mod	KORE50	0.2986	0.2986	0.2986	201.84
AGDISTIS	MSNBC	0.5277	0.5452	0.5113	1,176.8947
AGDISTIS-mod	MSNBC	0.5523	0.5523	0.5523	928.2
AGDISTIS	Microposts2014-Test	0.3177	0.3177	0.3177	82.24
AGDISTIS-mod	Microposts2014-Test	0.3209	0.3209	0.3209	76.1573
AGDISTIS	Microposts2014-Train	0.4186	0.4186	0.4186	108.6434
AGDISTIS-mod	Microposts2014-Train	0.4199	0.4199	0.4199	98.1222
AGDISTIS	N3-Reuters-128	0.6348	0.6366	0.633	482.7229
AGDISTIS-mod	N3-Reuters-128	0.6261	0.6261	0.6261	289.7656
AGDISTIS	N3-RSS-500	0.6016	0.6022	0.601	168.6733
AGDISTIS-mod	N3-RSS-500	0.614	0.614	0.614	126.31
AGDISTIS	OKE 2015 Task 1 evaluation dataset	0.5811	0.5825	0.5798	375.67
AGDISTIS-mod	OKE 2015 Task 1 evaluation dataset	0.5723	0.5723	0.5723	308.2475
AGDISTIS	OKE 2015 Task 1 example set	1	1	1	120.6667
AGDISTIS-mod	OKE 2015 Task 1 example set	1	1	1	155.3333
AGDISTIS	OKE 2015 Task 1 gold standard sample	0.6114	0.6227	0.6006	329.6489
AGDISTIS-mod	OKE 2015 Task 1 gold standard sample	0.6568	0.6568	0.6568	251.7708

AGDISTIS results

- 20 % time reduction
- Increased precision and recall

Conclusion

- Computing semantic relatedness in a scalable way is important
- For QA we also need to take care of the relations
- There is more research needed to find a good equilibrium between time and memory consumption

Thank you :)