Ranking vs. Classifying: Measuring Knowledge Base Completion Quality

Marina Speranskaya, Martin Schmitt, Benjamin Roth

CIS, LMU Munich

22.6.2020 – 24.6.2020

Motivation

Rank = 3, not bad!

Thomas Lennon – has profession – ?

top-3

e4	musician	0.9
e7	artist	0.8
e3	film producer	0.77
e2	screenwriter	0.5
	...	
e1	New Zealand	0.04
e6	US dollar	0.01

➡ Precision & recall

Top-3, top-5?

With what triples do I actually update my KB?

How does it change the **KB quality**?

TP = {e3}
FP = {e4, e7}
FN = {e2}

Motivation

Thomas Lennon – has professio

What other possible scenarios are there?

Reflects the closed-world problematic

Top-3

e4	musician	0.9		0.8	0.8
e7	artist	0.8		0.72	0.5
e3	film producer	0.77		0.4	0.3
e2	screenwriter	0.5		0.3	0.2
	...				
e1	New Zealand	0.04		0.04	
e6	US dollar	0.01		0.01	

Paris – has profession - ?

➡ Precision & recall

Motivation

How to account for different number of answers?

Thomas Lennon – has professi

top-3

e4	musician	0.9	0.8	0.8
e7	artist	0.8	0.72	0.5
e3	film producer	0.77	0.4	0.3
e2	screenwriter	0.5	0.3	0.2
	...			
e1	New Zealand	0.04		04
e6	US dollar	0.01		0.01

> 0.7

Despite good ranking the classification is not optimal yet..

➡ Precision & recall

➡ Score-based threshold

Motivation

How to account for different number of answers?

Thomas Lennon – has professi

e4	musician	0.9	
e7	artist	0.8	
e3	film producer	0.77	
e2	screenwriter	0.5	
	...		
e1	New Zealand	0.04	
e6	US dollar	0.01	

> 0.7

~~0.8~~	0.81	
~~0.72~~	0.75	
~~0.4~~	0.73	
~~0.3~~	0.71	
		~~0.04~~ 0.03
		~~0.01~~ 0.02

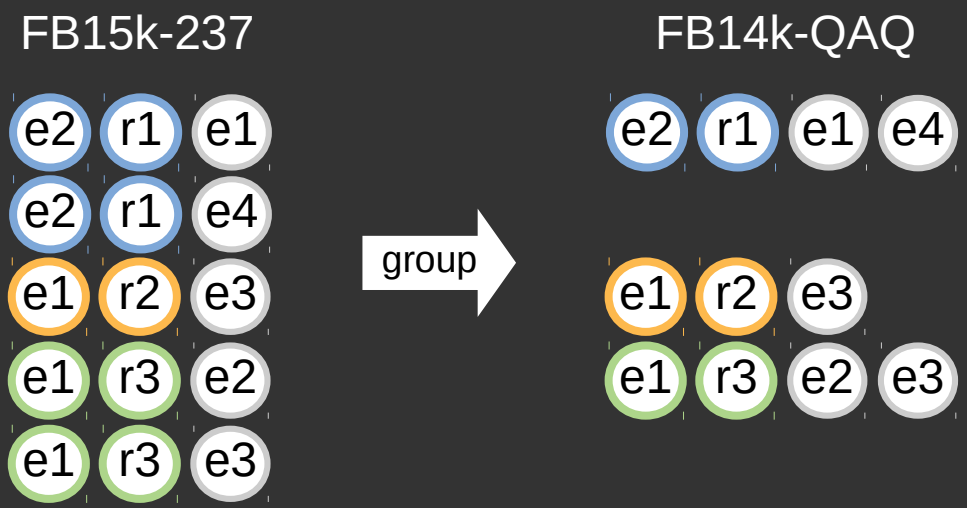
~~0.8~~	0.5
~~0.5~~	0.5
~~0.3~~	0.3
~~0.2~~	0.3

Despite good ranking the classification is not optimal yet..

➡ Precision & recall

➡ Score-based threshold ➡ Score calibration

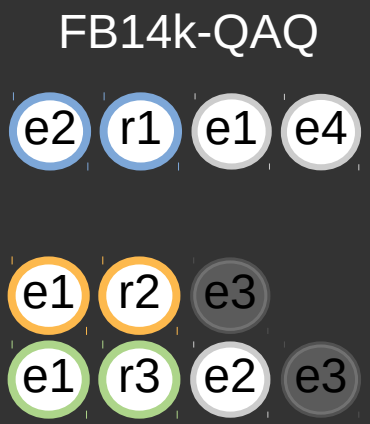
FB14k-QAQ: Query Answering Quality



- **Precision & recall**
- **Query-based**
- **Score-based threshold**
- **Score calibration**

FB14k-QAQ: Query Answering Quality

group



new evaluation set

e2, r1, {e1, e4}
e1, r2, {}
e1, r3, {e2}

+ nonsensical

Make inaccessible by removing

→ **Precision & recall** → **Query-based**

→ **Score-based threshold** → **Score calibration**

Results

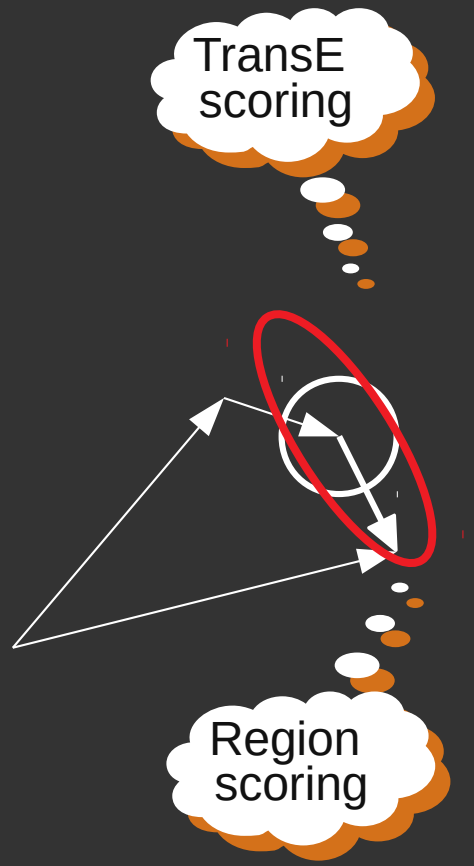
Model (d)	MRR		F1
ConvE 128	**.321**	1 3	.134
ConvE 64	.263	7 2	.157
ComplEx 128	.293	2 7	.021
ComplEx 64	.293	3 8	.009
TransE 128	.293	4 6	.108
TransE 64	.283	5 5	.111
DistMult 64	.266	6 1	**.159**
DistMult 128	.221	8 4	.133

Complete > nonsensical > inaccessible

Relative order of the models changes

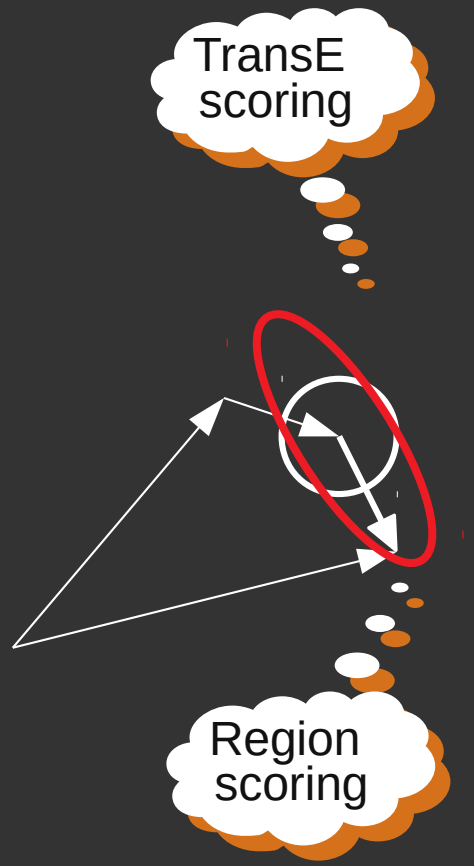
Results

Model (d)	MRR	F1
ConvE 128	**.321**	.134
ConvE 64	.263	.157
ComplEx 128	.293	.021
ComplEx 64	.293	.009
TransE 128	.293	.108
TransE 64	.283	.111
DistMult 64	.266	**.159**
DistMult 128	.221	.133



Results

Model (d)	MRR	F1
ConvE 128	**.321**	.134
ConvE 64	.263	.157
ComplEx 128	.293	.021
ComplEx 64	.293	.009
TransE 128	.293	.108
TransE 64	.283	.111
DistMult 64	.266	**.159**
DistMult 128	.221	.133
Region 64	.330	.146



Contributions

1. A classification-based evaluation setting that provides **interpretable performance metrics**

2. A new benchmark FB14k-QAQ reflecting a **comprehensive set of KBC scenarios**

3. A simple, yet effective strategy to improve TransE's **ability to calibrate scores**

Available at: https://github.com/marina-sp/classification_lp

Thank you for your attention!

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - RO 5127/2-1, and by the BMBF as part of the project MLWin (01IS18050).

Results

C – complete
F – nonsensical
I – incomplete

Different complexity of the sets

Model (d)	MRR		F_1 global threshold				F_1 multiple thresholds			
			full	$\mathcal{C}$	$\mathcal{C} \sqcup \mathcal{F}$	$\mathcal{I}$	full	$\mathcal{C}$	$\mathcal{C} \sqcup \mathcal{F}$	$\mathcal{I}$
ConvE 128	**.321**	1	.134	.272	.211	.105	**.204**	**.317**	**.286**	.150
ConvE 64	.263	7	.157	**.307**	.261	.108	.189	.312	.280	.135
ComplEx 128	.293	2	.021	.169	.017	.042	.190	.296	.261	.143
ComplEx 64	.293	3	.009	.157	.005	.057	.181	.282	.245	.143
TransE 128	.293	4	.108	.156	.106	.108	.159	.172	.168	**.154**
TransE 64	.283	5	.111	.164	.112	**.110**	.161	.176	.175	**.154**
DistMult 64	.266	6	**.159**	.273	**.226**	.129	.184	.256	.239	.148
DistMult 128	.221	8	.133	.275	.[illegible]	.[illegible]	.206	.194	.138	

Relative order of the models changes