A Study on Entity Linking Across Domains: Which Data is Best for Fine-Tuning?

Hassan Soliman[1,2], Heike Adel[1], Mohamed Gad-Elrab[1], Dragan Milchevski[1], Jannik Strötgen[1]
[1]Bosch Center for Artificial Intelligence, Germany
[2]Saarland University, Germany

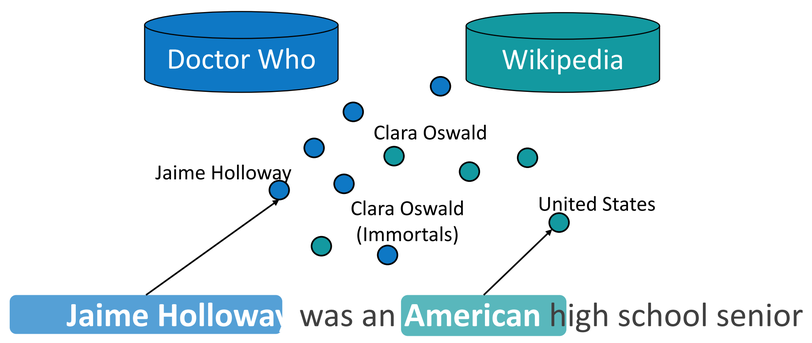
1 Cross-Domain Entity Linking

Domain-specific documents often contain mentions of entities from different knowledge graphs (KGs):
- General-domain KGs, e.g., Wikipedia (here: source domain)
- Domain-specific KGs, e.g., Fandom Wiki (here: target domain)

⇒ Required: Entity linking systems that can link to several KGs at the same time

Approach: joint vector space for different KGs

Research questions:
- Which data is best suited for fine-tuning?
- Does fine-tuning on domain-specific data harm performance on the general domain?

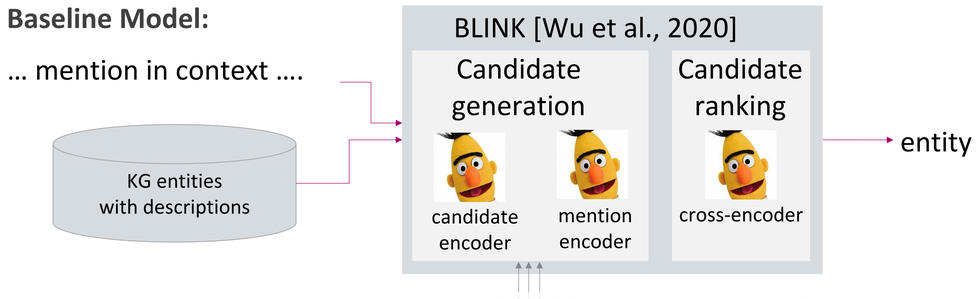


Contributions:
- Analysis of different data sources for fine-tuning a joint vector space
- Publication of list of overlapping entities (entities appearing in both general-domain KG and domain-specific KG)

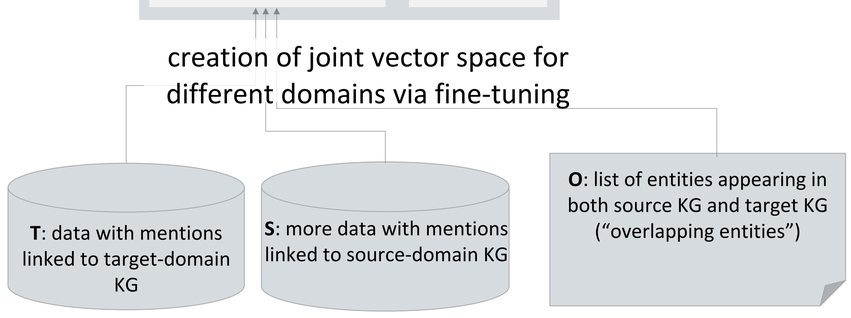
2 Model and Datasets

Extended Entity Linking Model

Baseline Model:



Extension to New Domains:



Fine-tuning the model using different data configurations: S, T, TO, TS, TOS

Data

- Zeshel dataset [Logeswaran et al., 2019]:
  focus on four target domains:
  American Football, Doctor Who, Fallout, Final Fantasy
- Reddit dataset [Botzer et al., 2021]:
  mentions annotated with entities from the source-domain KG (Wikipedia)

Domain	Entities	Fine-tuning Mentions			Overlapping Entities
		Train	Dev	Test	
American Football	31,929	3,000	320	578	22,928
Doctor Who	40,281	6,360	640	1,334	3,611
Fallout	16,992	2,500	320	466	752
Final Fantasy	14,044	4,360	640	1,041	413
Wikipedia (Reddit)	5,903,538	7,711	409	1,328	-

3 Results

Proximity of overlapping entities in joint vector space

Intuition: The closer the overlapping entities, the better the representation
Metrics: Mean Reciprocal Rank (MRR); Average Cosine Similarity (ACS)

Target KB	American Football		Doctor Who		Fallout		Final Fantasy	
Model	MRR	ACS	MRR	ACS	MRR	ACS	MRR	ACS
BLINK	0.4991	0.9938	0.4607	0.9650	0.4071	0.9603	0.3623	0.9532
T	0.4982	0.9892	0.3926	0.9095	0.3533	0.9317	0.4136*	0.9515
TO	0.4990	0.9919	**0.4932***	0.9784*	**0.4558***	0.9680*	**0.4400***	0.9628
TS	**0.4999**	**0.9958***	0.4323	0.9605	0.4223*	0.9676*	0.4072*	0.9746*
TOS	0.4995	0.9896	0.4619	**0.9830***	0.4534*	**0.9820***	0.4209*	**0.9791***

Observations:
- Fine-tuning on target data only (T) is not sufficient
- Fine-tuning on overlapping entities (O) improves the vector space

Entity Linking on Target-Domain KG

Observation: Using source-domain data for fine-tuning even helps for entity linking on the target domain

Target KG	American Football		Doctor Who		Fallout		Final Fantasy	
Model	AP@1	MAP@10	AP@1	MAP@10	AP@1	MAP@10	AP@1	MAP@10
BLINK	0.1747	0.4104	0.4108	0.4810	0.3412	0.4444	0.3833	0.5179
S	0.1713	0.3732	0.5337*	0.6191*	0.4249*	0.5295*	0.3881	0.5433
T	**0.2093***	**0.4606***	**0.6169***	0.6925*	0.4313*	0.5510*	0.3871	0.5405
TO	0.1938	0.4103	0.5697*	0.6558*	0.4485*	0.5590*	0.3439	0.4881
TS	0.2076	0.4583*	0.6124*	**0.7124***	**0.4657***	**0.5915***	**0.4121**	**0.5710***
TOS	0.1540	0.3292	0.5345*	0.6149*	0.4227*	0.5405*	0.3910	0.5486*

Entity Linking on Source-Domain KG

Observation: Including source-domain data is key to avoid performance losses on source domain after fine-tuning

Target KG	American Football		Doctor Who		Fallout		Final Fantasy	
Model	AP@1	MAP@10	AP@1	MAP@10	AP@1	MAP@10	AP@1	MAP@10
BLINK	0.8479	0.8973	0.8509	0.8985	0.8509	0.8987	0.8494	0.8987
S	**0.8727**	**0.9051**	0.8750*	**0.9063**	**0.8788***	**0.9089**	0.8758*	0.9070
T	0.8539	0.8994	0.8209	0.8556	0.8057	0.8464	0.8599	0.8991
TO	0.8524	0.8953	0.8532	0.8865	0.8381	0.8714	0.8630	0.8956
TS	0.8607	0.8957	0.8582	0.8976	0.8599	0.8965	0.8788*	**0.9085**
TOS	0.8170	0.8386	**0.8773***	0.9062	0.8637	0.8858	**0.8795***	0.9038

* denotes statistical significant differences to BLINK model (randomization test, α = 0.005 with Bonferroni correction)