\Orchestrating a brighter world NEC

QA-Matcher: Unsupervised Entity Matching Using A Question Answering Model

Shogo Hayashi, Yuyang Dong, Masafumi Oyamada

Entity matching: background

- Entity matching (EM) is to match the records between two tables that refer to the same real-word entity. It is a fundamental problem in data integration.

ID	title	manufacturer	price
L1	motu digital performer 5 digital audio software competitive upgrade (mac only)	motu	395.0
L2	illustrator cs3 13 mac ed 1u	adobe-education-box	199.0
L3	microsoft visio standard 2007 version upgrade	microsoft	129.95

Amazon

matching

ID	title	manufacturer	price
R1	motu digital performer dp5 software music production software	NaN	319.95
R2	adobe illustrator cs3 for mac academic	adobe-education-box	199.99
R3	adobe cs3 design standard upgrade	NaN	413.99

Google

- A popular framework is blocking and matching, we focus on matching
  - Blocking: Quick generate candidate pairs from two datasets
  - Matching: classify the candidate pairs into “match” or “non-match”
    - e.g., M(L1, R1) -> 1 or 0

Entity matching: related works

- Supervised learning approach
  - Build a model with human labelled data
  - SoTA [Ditto VLDB21]: Serialize the row values to text and then do text classification with PLM (BERT)

- Unsupervised learning approach
  - Build a model without human intervention
  - SoTA [ZeroER SIGMOD20]: Generate similarity values as feature vectors and cluster the vectors into “match” and “non-match” clusters using GMM.

- Zero-shot approach
  - No training, no labelled data, just adjust other trained model
  - prompt language model to solve entity matching task -> This work

Question answering: background

- Given a **question** and a **passage**, output the **answer**

Question:
What's the name of the software used to manage music and other media on Apple devices?

Passage:
Apple's iTunes software (and other alternative software) can be used to transfer music, photos, videos, games, contact information, e-mail settings, Web bookmarks, and calendars, to the devices supporting these features from computers using certain versions of Apple Macintosh and Microsoft Windows operating systems.

Answer:
iTunes

- Many QA models (LM+ finetune on QA datasets, e.g.) are available.

Idea: solve entity matching as question answering

ID	title	manufacturer	price
L1	motu digital performer 5 digital audio software competitive upgrade (mac only)	motu	395.0
L2	illustrator cs3 13 mac ed 1u	adobe-education-box	199.0
L3	microsoft visio standard 2007 version upgrade	microsoft	129.95

Amazon

matching

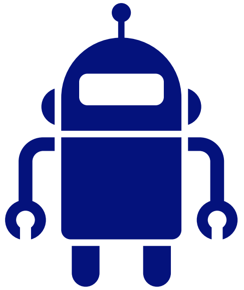
ID	title	manufacturer	price
R1	motu digital performer dp5 software music production software	NaN	319.95
R2	adobe illustrator cs3 for mac academic	adobe-education-box	199.99
R3	adobe cs3 design standard upgrade	NaN	413.99

Google

Question
What is characterized by motu digital performer 5 digital audio software competitive upgrade (mac only) motu 395.0?

Passage:
R1 is characterized by motu digital performer dp5 software music production software 319.95.
R2 is characterized by adobe illustrator cs3 for mac academic 199.99.
R3 is characterized by adobe cs3 design standard upgrade 413.99.

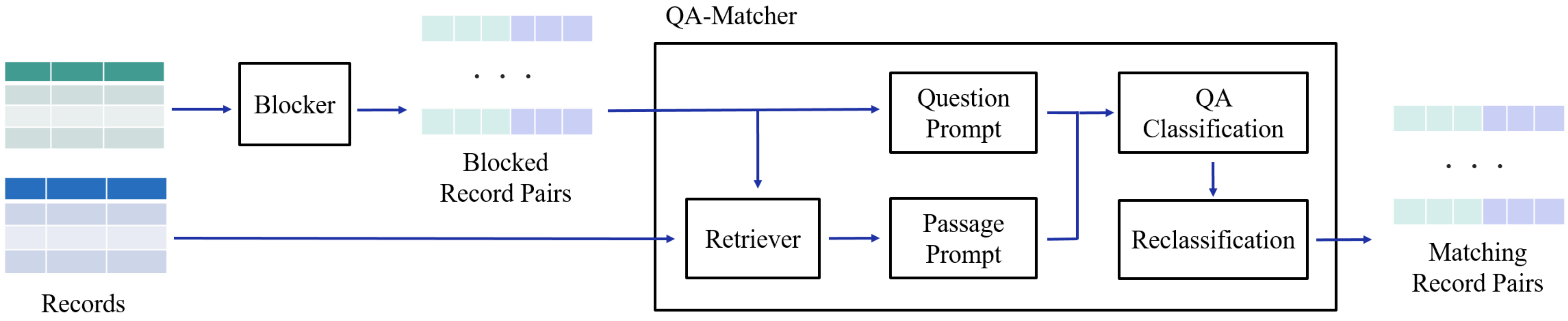
Answer:
R1.



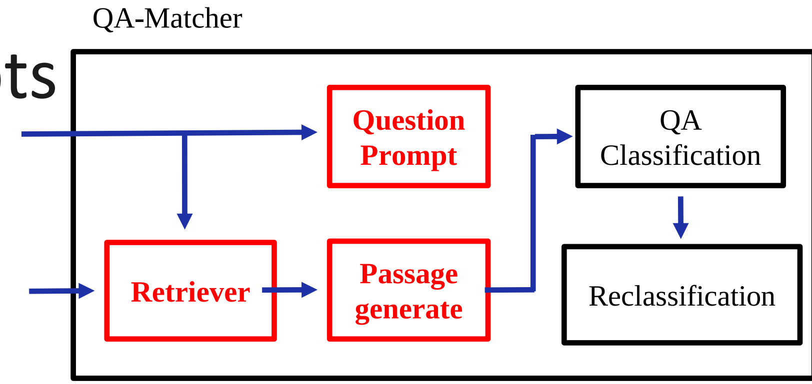
QA Model

Overall of entity matching processing with QA-matcher

- Steps
  - Retrieve relevant records, question and passage prompts
  - QA classification
  - Reclassification



Retrieve relevant records, question and passage prompts



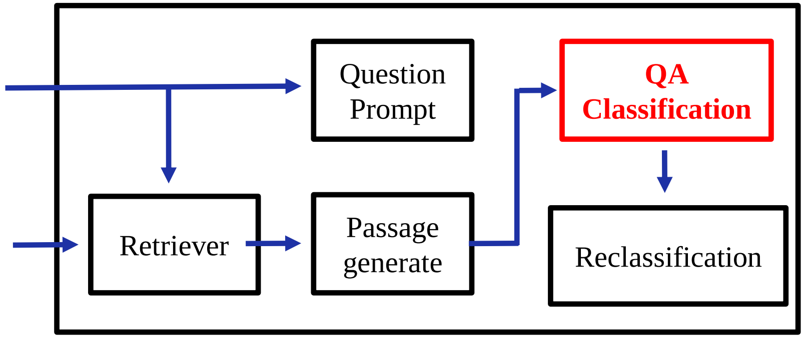
- Given a record pair <l1, r1> and two datasets L and R
  - For l1
    - Retrieve top-k relevant records from R with nearest neighbor search
    - Question: q(l1) = “What is characterized by $l1.values$”
    - Passage: p(l1) = “ $r1.id$ is characterized by $r1.values$. …, rk is characterized by $rk.values$, ”
  - For r1
    - same way… and got q(r1) and p(r1)

Question (L1):
What is characterized by motu digital performer 5 digital audio software competitive upgrade (mac only) motu 395.0?

Passage:
R1 is characterized by motu digital performer dp5 software music production software 319.95.
R2 is characterized by adobe illustrator cs3 for mac academic 199.99.
R3 is characterized by adobe cs3 design standard upgrade 413.99.

QA Classification

QA-Matcher



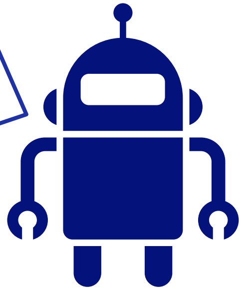
- Then ask QA model with q(l1), p(l1) and q(r1)·p(r1)
- If the answer is just same of given pair (r1, l1), they are matching.

.

Question q(l1):
What is characterized by motu digital performer 5 digital audio software competitive upgrade (mac only) motu 395.0?

Passage p(l1):
R1 is characterized by motu digital performer dp5 software music production software 319.95.
R2 is characterized by adobe illustrator cs3 for mac academic 199.99.
R3 is characterized by adobe cs3 design standard upgrade 413.99.

Answer:
r1.



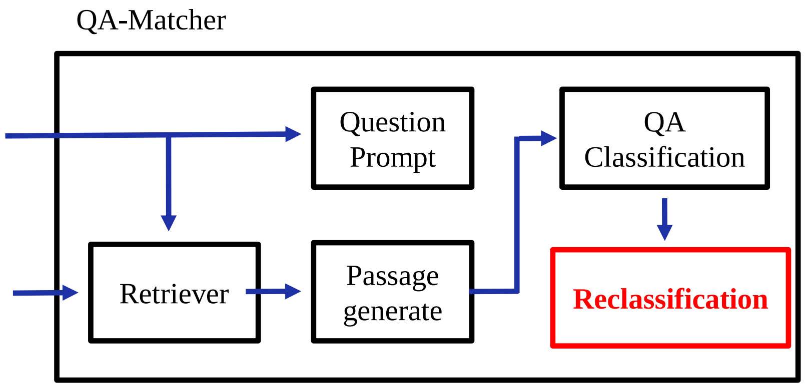
QA Model

Question q(r1):
What is characterized by motu digital performer dp5 software music production software 319.95?

Passage p(r1):
L1 is characterized by motu digital performer 5 digital audio software competitive upgrade (mac only) motu 395.0.
L2 is characterized by illustrator cs3 13 mac ed 1u adobe-education-box 199.0.
L3 is characterized microsoft visio standard 2007 version upgrade Microsoft 395.0.

Answer:
l1.

Reclassification



- To calibrate the result and get further performance improvement, we conduct a reclassification step.
- Score = 1/2 * ($\underbrace{\text{score\_QA(r1 | q(l1), p(l1))}}_{\text{score of QA answer "r1"}}$ + $\underbrace{\text{score\_QA(l1 | q(r1), p(r1))}}_{\text{score of QA answer "l1"}}$)

- Train a one-dimensional linear classifier with {score, QA_predict} pairs with maximizing likelihood.
- False positives (false negatives) are expected to have low (high) scores, and the reclassification can correct them.

Experiments: setting

- Datasets:
  - We use 16 datasets with structured, dirty, textual, and heterogeneous types collected from existing works and benchmarks.
- Comparing Methods
  - ZeroER [SIGMOD20]
  - Ditto [VLDB21]
  - Deepmatcher [SIGMOD18]
- Configuration of QA-matcher
  - QA model: Bert-large-finetuned-squad
  - Related records k=20

Table 1: Summary of benchmark datasets.

Type	Dataset	Domain	Size	# Attr.
Structured (two record sets share the same schema or attributes)	BeerAdvo-RateBeer	beer	91	4
	iTunes-Amazon$_1$	music	109	8
	Fodors-Zagats	restaurant	189	6
	DBLP-ACM$_1$	citation	2473	4
	DBLP-Scholar$_1$	citation	5742	4
	Amazon-Google	software	2293	3
	Walmart-Amazon$_1$	electronics	2049	5
Dirty (some attribute values are injected in wrong attributes)	iTunes-Amazon$_2$	music	109	8
	DBLP-ACM$_2$	citation	2473	4
	DBLP-Scholar$_2$	citation	5742	4
	Walmart-Amazon$_2$	electronics	2049	5
Textual (attribute values are long texts)	Abt-Buy	product	1916	3
	Company	company	22503	1
Heterogeneous (two record sets do not share the same schema)	Walmart-Amazon$_3$	electronics	2049	(4, 5)
	Walmart-Amazon$_4$	electronics	2049	(4, 4)
	Walmart-Amazon$_5$	electronics	2049	(4, 4)

Experiments: QA-Matcher vs ZeroER:

- QA-Matcher strongly outperforms ZeroER on 15/16 datasets.

Dataset	Unsupervised					Supervised	
	QA-Matcher	Sentence-BERT (QA-Matcher w/o QA)	QA-Matcher w/o retriever	QA-Matcher w/o reclas.	ZeroER	Ditto	Deep Matcher
BeerAdvo-RateBeer	**0.9333**	0.8966	0.4667	**0.9333**	0.7407	0.9437	0.7880
DBLP-Scholar$_1$	<u>0.7276</u>	0.5758	0.3355	0.6671	**0.7921**	0.9560	0.9470
DBLP-ACM$_1$	0.9754	<u>0.9865</u>	0.3844	**0.9888**	0.9541	0.9899	0.9845
Fodors-Zagats	**1.0000**	0.9333	0.2178	0.7586	<u>0.9767</u>	1.0000	1.0000
iTunes-Amazon$_1$	**0.9434**	0.6977	0.3971	<u>0.8302</u>	0.4800	0.9706	0.9120
Amazon-Google	<u>0.6555</u>	0.5266	0.2285	**0.6992**	0.2027	0.7558	0.7070
Walmart-Amazon$_1$	**0.6907**	0.4344	0.1722	0.6064	<u>0.6645</u>	0.8676	0.7360
DBLP-ACM$_2$	0.9832	<u>0.9853</u>	0.3771	**0.9921**	0.3974	0.9903	0.9810
iTunes-Amazon$_2$	**0.7727**	0.2500	0.4091	<u>0.6957</u>	0.4333	0.9565	0.7940
DBLP-Scholar$_2$	**0.7342**	0.5748	0.3379	<u>0.6635</u>	0.4169	0.9575	0.9380
Walmart-Amazon$_2$	**0.6433**	0.4855	0.1722	<u>0.5808</u>	0.2329	0.8569	0.5380
Abt-Buy	<u>0.7218</u>	0.3828	0.2205	**0.7569**	0.2110	0.8933	0.6280
Company	<u>0.5197</u>	0.4554	0.4308	**0.6131**	–	0.9385	0.9270
Walmart-Amazon$_3$	**0.6653**	0.4329	0.1722	<u>0.5803</u>	0.0000	0.8106	0.6710
Walmart-Amazon$_4$	**0.6653**	0.4329	0.1723	<u>0.5803</u>	0.0000	0.8211	0.6340
Walmart-Amazon$_5$	**0.6565**	0.4494	0.1725	<u>0.5556</u>	0.0000	0.8140	0.6650

Experiments: unsupervised vs supervised

- QA-Matcher (zero-shot) it is competitive with supervised learning SoTA methods.

Dataset	Unsupervised					Supervised	
	QA-Matcher	Sentence-BERT (QA-Matcher w/o QA)	QA-Matcher w/o retriever	QA-Matcher w/o reclas.	ZeroER	Ditto	Deep Matcher
BeerAdvo-RateBeer	**0.9333**	0.8966	0.4667	**0.9333**	0.7407	0.9437	0.7880
DBLP-Scholar$_1$	0.7276	0.5758	0.3355	0.6671	**0.7921**	0.9560	0.9470
DBLP-ACM$_1$	0.9754	0.9865	0.3844	**0.9888**	0.9541	0.9899	0.9845
Fodors-Zagats	**1.0000**	0.9333	0.2178	0.7586	0.9767	1.0000	1.0000
iTunes-Amazon$_1$	**0.9434**	0.6977	0.3971	0.8302	0.4800	0.9706	0.9120
Amazon-Google	0.6555	0.5266	0.2285	**0.6992**	0.2027	0.7558	0.7070
Walmart-Amazon$_1$	**0.6907**	0.4344	0.1722	0.6064	0.6645	0.8676	0.7360
DBLP-ACM$_2$	0.9832	0.9853	0.3771	**0.9921**	0.3974	0.9903	0.9810
iTunes-Amazon$_2$	**0.7727**	0.2500	0.4091	0.6957	0.4333	0.9565	0.7940
DBLP-Scholar$_2$	**0.7342**	0.5748	0.3379	0.6635	0.4169	0.9575	0.9380
Walmart-Amazon$_2$	**0.6433**	0.4855	0.1722	0.5808	0.2329	0.8569	0.5380
Abt-Buy	0.7218	0.3828	0.2205	**0.7569**	0.2110	0.8933	0.6280
Company	0.5197	0.4554	0.4308	**0.6131**	–	0.9385	0.9270
Walmart-Amazon$_3$	**0.6653**	0.4329	0.1722	0.5803	0.0000	0.8106	0.6710
Walmart-Amazon$_4$	**0.6653**	0.4329	0.1723	0.5803	0.0000	0.8211	0.6340
Walmart-Amazon$_5$	**0.6565**	0.4494	0.1725	0.5556	0.0000	0.8140	0.6650

Experiments: ablation study

- QA is better than cross-encoder
- QA with random retriever is poor.
- Reclassification significantly improved the F1 score for 10 datasets, but it slightly decreased the score for the remaining 6 datasets.

Dataset	Unsupervised			
	QA-Matcher	Sentence-BERT (QA-Matcher w/o QA)	QA-Matcher w/o retriever	QA-Matcher w/o reclas.
BeerAdvo-RateBeer	**0.9333**	0.8966	0.4667	**0.9333**
DBLP-Scholar$_1$	<u>0.7276</u>	0.5758	0.3355	0.6671
DBLP-ACM$_1$	0.9754	<u>0.9865</u>	0.3844	**0.9888**
Fodors-Zagats	**1.0000**	0.9333	0.2178	0.7586
iTunes-Amazon$_1$	**0.9434**	0.6977	0.3971	<u>0.8302</u>
Amazon-Google	<u>0.6555</u>	0.5266	0.2285	**0.6992**
Walmart-Amazon$_1$	**0.6907**	0.4344	0.1722	0.6064
DBLP-ACM$_2$	0.9832	<u>0.9853</u>	0.3771	**0.9921**
iTunes-Amazon$_2$	**0.7727**	0.2500	0.4091	<u>0.6957</u>
DBLP-Scholar$_2$	**0.7342**	0.5748	0.3379	<u>0.6635</u>
Walmart-Amazon$_2$	**0.6433**	0.4855	0.1722	<u>0.5808</u>
Abt-Buy	<u>0.7218</u>	0.3828	0.2205	**0.7569**
Company	<u>0.5197</u>	0.4554	0.4308	**0.6131**
Walmart-Amazon$_3$	**0.6653**	0.4329	0.1722	<u>0.5803</u>
Walmart-Amazon$_4$	**0.6653**	0.4329	0.1723	<u>0.5803</u>
Walmart-Amazon$_5$	**0.6565**	0.4494	0.1725	<u>0.5556</u>

$\underbrace{\hspace{6em}}_{\text{Ablation study}}$

Experiments: effect on retrieved number k

- Not significantly affected by k.
- Nearest neighbor retriever can retrieve candidates correctly.

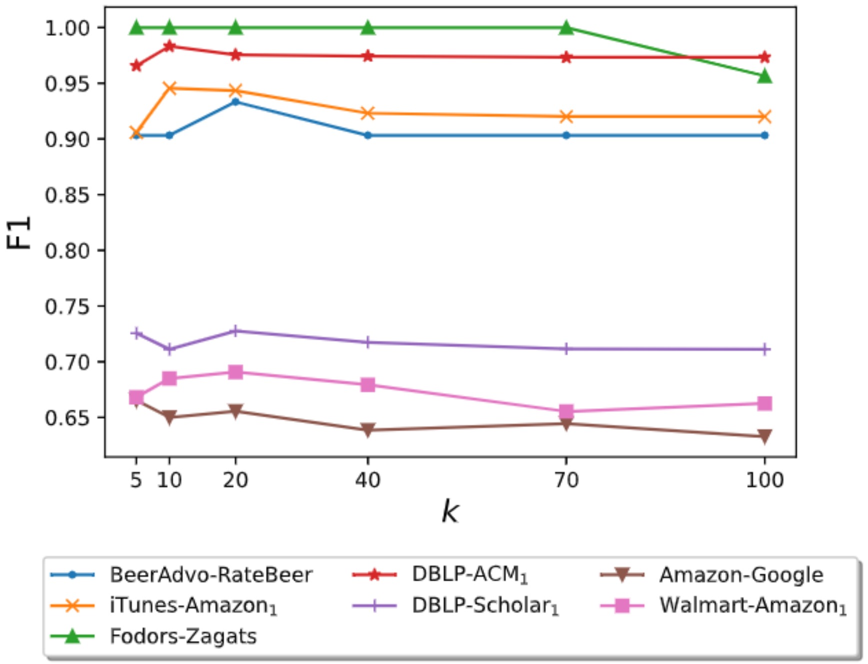


Fig. 5: F1 scores for different numbers of retrieved records k.

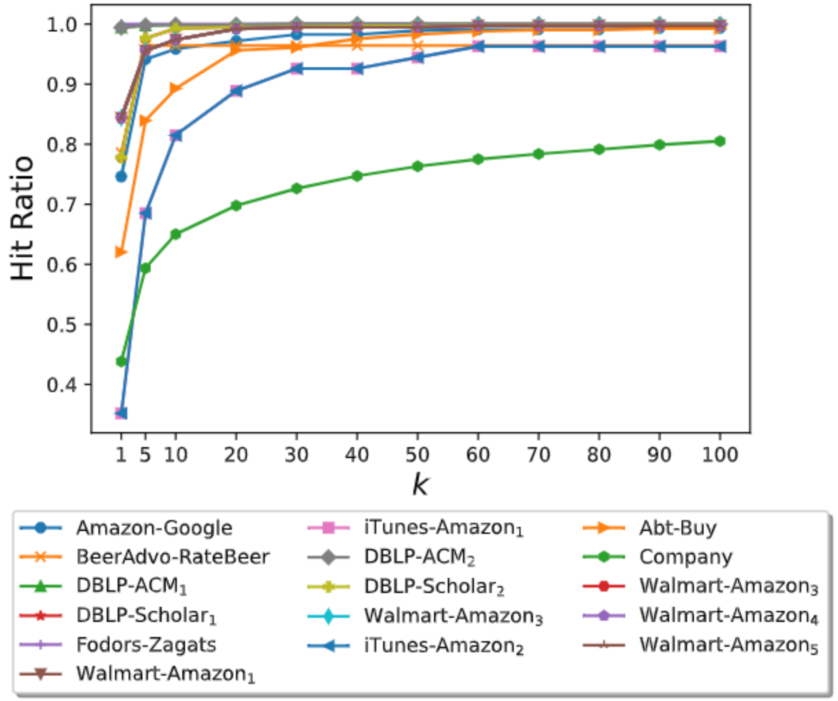


Fig. 6: Hit ratio of the retriever for different numbers of retrieved records k.

Summary and Thanks!