

QUOREF: A Reading Comprehension Benchmark with Questions Requiring Coreferential Reasoning

Pradeep Dasigi (AI2), Nelson F. Liu (S), Ana Marasović (AI2), Noah A. Smith (W, AI2), Matt Gardner (AI2)



24K span-selection questions on English Wikipedia passages

Questions require coreference resolution (est. 78%)

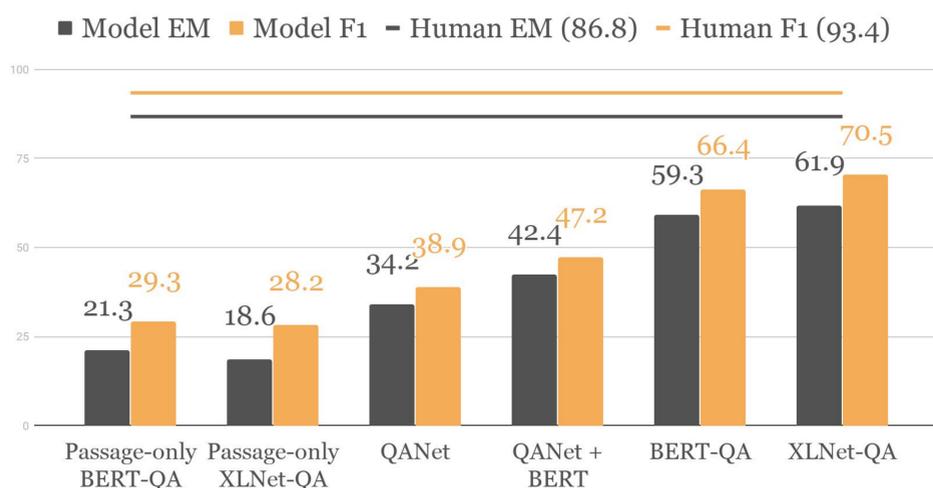
Answers with multiple spans (10%)

Built with strong adversary in loop

Pronoun resolution* (69%)	Anna and Declan [...] discover the three van thieves going through Anna's luggage. Declan fights them , [...]	Who does Declan get into fight with?	the three van thieves
Nominal resolution* (54%)	Aside from Constantinople and Trebizond, other Byzantine cities also featured tzykanisteria , most notably Sparta, Ephesus, and Athens, [...]	Which cities had tzykanisteria?	Constantinople; Trebizond; Sparta; Ephesus; Athens
Multiple resolutions* (32%)	The now upbeat collective keep the toucan , nicknaming it " Amigo " [...] When authorities show up to catch the bird , Pete and Liz spirit him away by Liz hiding him in her dress	What is the name of the character who hides in Liz's dress?	Amigo
Common-sense* (10%)	Amos [...] with his mother Fania and father Arieh . One of his mother's friends is killed while hanging up laundry during the war[...] Fania fall into depression [...] kills herself by overdose [...]	How does Arieh's wife die?	kills herself by overdose

highlight: question anchor; **bold**: relevant coreference

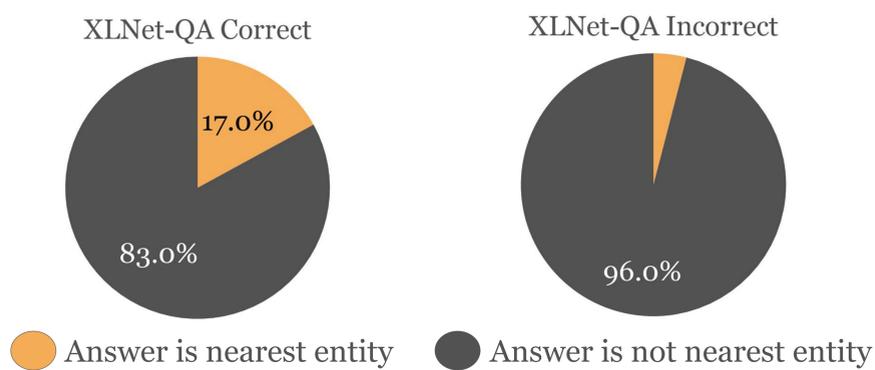
* estimated from 100 training instances



Baseline Performance

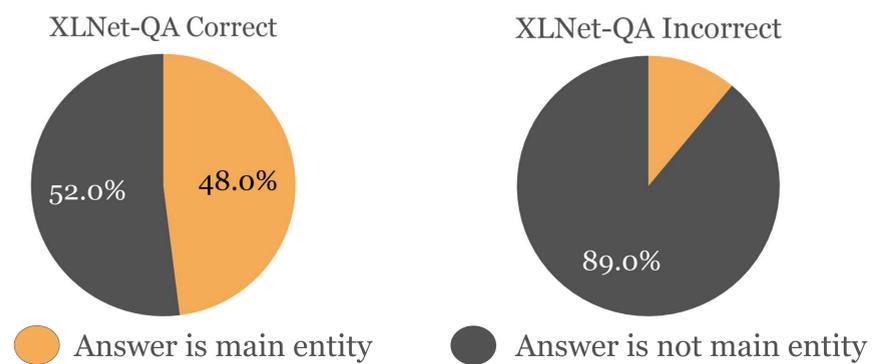
- Passage-only models show extent of artifacts
- Best model << human performance (**23 F1 points**)
- Human performance est. from 400 test instances

Error Analysis: Distance



- Manually estimated from 100 dev instances
- XLNet-QA exploits entities being close to question anchors

Error Analysis: Main Entity



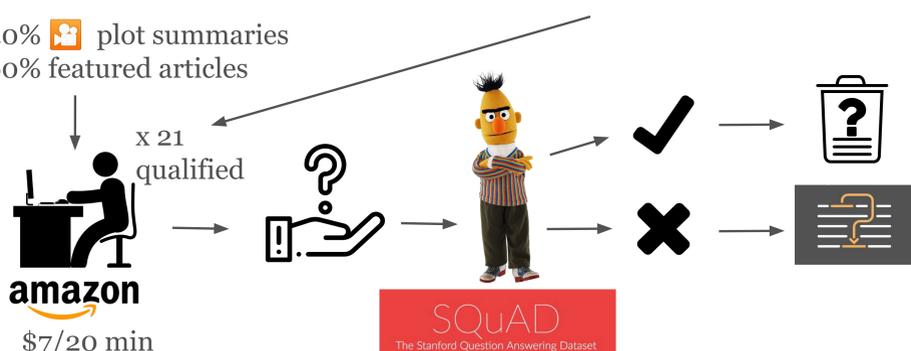
- *Main entity* is the first or the most frequent entity in passage ⇒ **artifact** of Wikipedia passages
- XLNet-QA exploits answers being main entities

Crowdsourcing Details



- Write questions that link multiple mentions of the same entity
- Follow good and bad examples

40% plot summaries
60% featured articles



Dataset Stats

	Train	Dev	Test
Questions	19399	2418	2537
Paragraphs	3771	454	477
Paragraph vocab	57648	18226	18885
Question vocab	19803	5579	5624
% multi-span answers	10.2	9.1	9.7

Dataset and leaderboard @ allennlp.org/quoref