Google Research

TyDi QA

A Benchmark for Information-Seeking **Q**uestion **A**nswering
in ***Ty***pologically ***Di***verse Languages

Jon Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Jennimaria Palomaki, Vitaly Nikolaev

Agenda

Goals

Challenge for ML

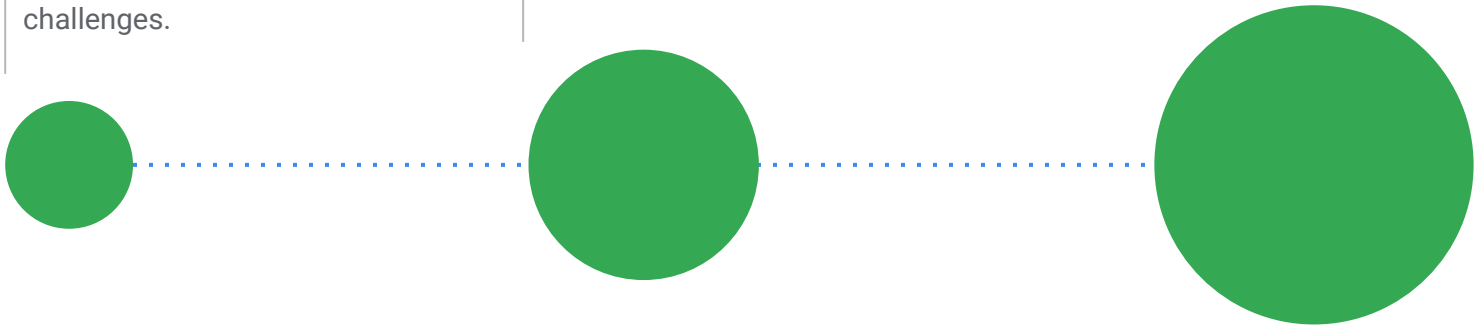
Non-English languages and less data leads to compelling challenges.

Opportunity for Science

Discover how the variety of human languages encode meaning.

NLP for everyone

Not everyone speaks English.



Your dataset determines your conclusions.

Affect your conclusions how?

SQuAD

Natural Questions

Open-Retrieval Question Answering (ORQA)

(Lee, 2019)

Scrap it.

-1.6 F1

Keep it.

+6.5 F1

Reading Comprehension vs. Information-Seeking QA

Passage

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called “showers”.

Annotator writes question

Question
What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

Answer
graupel

Question: What ship did Han Solo pilot?

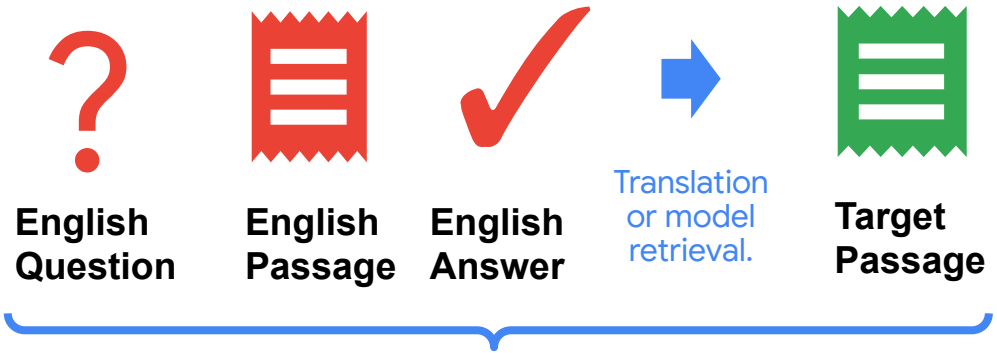
Annotator finds answer in article

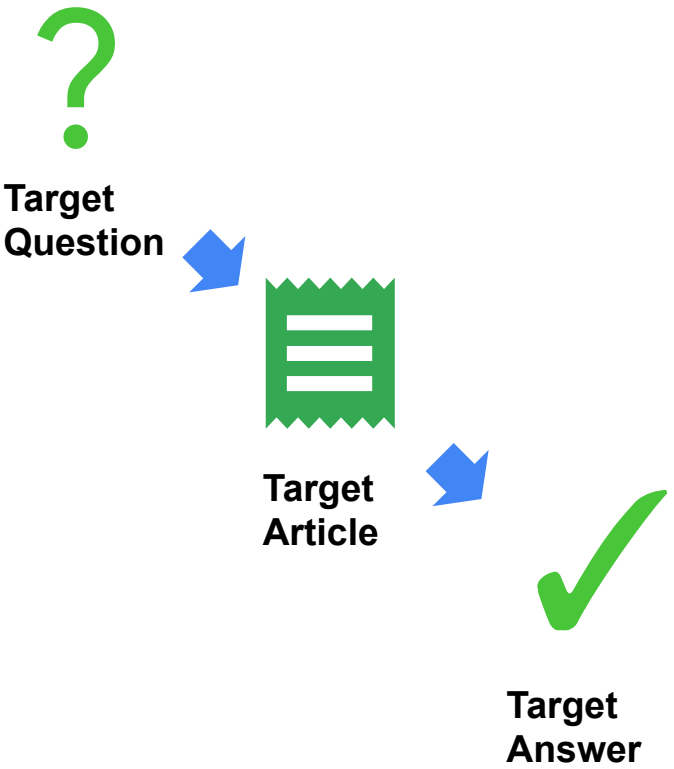
Article

The **Millennium Falcon** is a fictional starship in the Star Wars franchise. The modified YT-1300 Corellian light freighter is primarily commanded by Corellian smuggler Han Solo (Harrison Ford) and his Wookiee first mate, Chewbacca (Peter Mayhew). Designed by the Corellian Engineering Corporation (CEC), the highly modified YT-1300 is durable, modular, and is stated as being the second-fastest vessel in the Star Wars canon.

… and multiple choice, abstractive, cloze...

Going Multilingual: Translation vs. Direct Labeling





Does your **dataset** allow you to answer your **research questions**?

Research Questions

Does my multilingual model accurately represent human languages?

...or just English and its friends?

Typological Diversity

Diversity of **Available Data Types**

Does my multilingual model represent language as users use it?

...or some artificially clean version such as **Translationese**?

Elicitation, not translation

Information ***Seeking***

Does my model easily transfer to new languages?

...or are we stuck collecting data in 7000 languages?

Agenda

Typologically Diverse Languages

English	Fascinating language! ...but *no credit!*
Arabic	كتب
Bengali	সফেদা ফল খেতে কেমন
Finnish	jälleenrakennustöihin
Kiswahili	inayozungumzwa katika
Korean	우리 모두가 만들어가는
Indonesian	kecilkecilan
Japanese	24時間でのサーキット周回数
Russian	микроскопический
Telugu	ఖండాలలో అతిపెద్ద
Thai	เรือยิงตอร์ปิโดยูโกสลาเวีย ที5

Data Collection

Given prompt, write a question...

You're actually interested in.

Not answered by the prompt.

Read the question.

Locate an answer passage.

Highlight the minimal answer span.

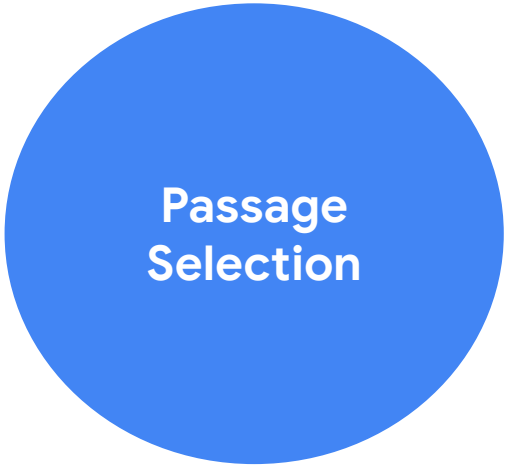
Question fluent?

Dialect/typos are fine!

Answer plausible?

100% in-language.

Primary Tasks



Given list of passages in article, select the answer passage (or NULL).

Given text of article, select the single contiguous span of bytes that is the answer (or NULL).

Gold Passage Task (SQuAD 1.1-like)

Show me the data already!

Kuka	keksi	viiko-n-päivät?
who	invented	week-**GEN**-day-**PL**

“Who invented the days of the week?”

Seitsen-päivä-inen	viikko	on	todennäköisesti	lähtöisin	Babylonia-sta
seven-NOM-day-PL.adj	week-NOM	is	likely	origin	Babylonia-**ELA**

“The seven-day week is most likely from Babylonia”

Challenge:

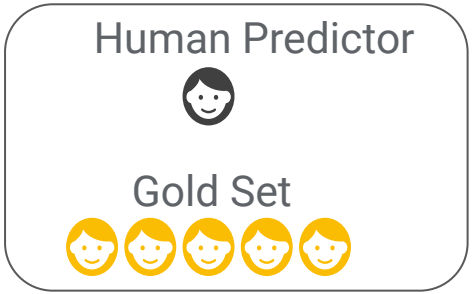
Finnish has heavy morphology and compounding, making question-answer matching hard (above).
Even the entity answer itself is inflected.

Agenda

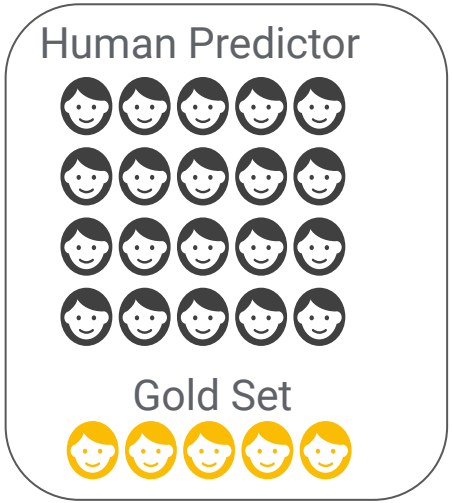
What even is “Human Performance”?

Not an upper bound for ML systems.

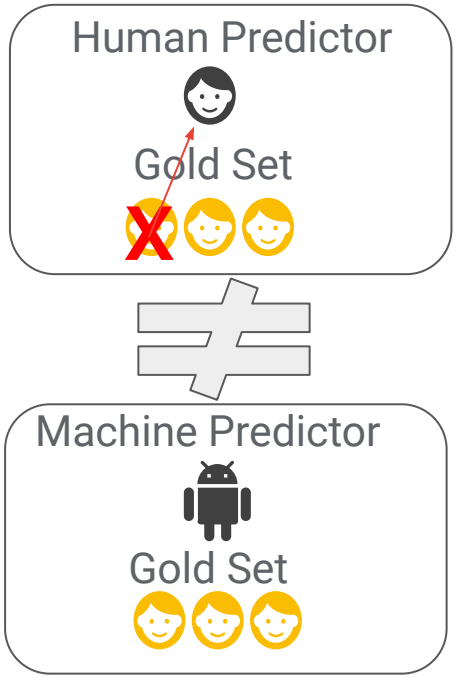
Single annotator



Super annotator



Pessimistic Human



Natural Questions	F1
Single annotator	57.5
Super annotator	75.5

Better estimate is +18 F1.

Baseline results

Caveat: Shouldn't compare across languages.

	Passage Answer			Minimal Answer Span	
Language	1st Passage	mBERT	Pessimistic Human	mBERT	Pessimistic Human
English	32.9	62.5	70.6	44.0	53.0
Indonesian	32.6	61.4	77.0	51.3	66.1
Russian	30.0	63.2	87.1	45.8	68.9
+9 more languages...					
Overall *(no English!)*	**30.2**	**63.1**	**79.9**	**50.5**	**70.1**

Short Term: Lots of room to improve toward Lesser Human.

Long Term: We can do better than Lesser Human.

What's in a leaderboard?

Is there a research paper describing this system?

Should tell us *why* and *how* it's better.

Is the source code available?

Replication is a key step if we're doing science.

Is the system trained on any additional public data?

We need to know what data to compare against.

Is the system trained on any additional private data?

That would mean we can't reproduce this system.

Is the system trained using any public APIs or private tools?

These are inherently not reproducible.

Agenda

Spelling Variation in Arabic Transliteration

Q: ؟ موزارت هو من
? mwzArt hw mn
Who is Mozart ?

A: بالنمسا سالزبورغ في 1756 يناير 27 في ولد (1791 ديسمبر 5 - 1756 يناير 27) موتسارت أماديوس فولفغانغ
bAlnmsA sAlzbwrg fy 1756 ynAyr 27 fy wld (1791 dysmbr 5 - 1756 ynAyr 27) mwtsArt A#mAdyws fwlfgAng
Wolfgang Amadeus Mozart (January 27, 1756 - December 5, 1791) was born on January 27, 1756 in Salzburg, Austria

Challenge: Same entity is spelled different ways.

Script Switching in Russian

Q:	Кто	изобрел	телефон	?			
	Kto	izobrel	telefon	?			
	who	invented	telephone	?			
	Who invented the		*telephone*	*?*			
A:	Сам	Рейс	назвал	сконструированное	им	устройство	Telephone .
	Sam	Reis	nazyval	skonstruirovannoe	im	ustroistvo	Telephone .
	self	Reis	called	constructed	him	device	Telephone .
	Reis himself called the device he created the						*Telephone .*

Challenge: Key concept is spelled using two different alphabets.

Vowel Diacritization in Arabic

Q: ؟ العُمانى العلم الوان هي ما
? AlEumAny AlElm AlwAn hy mA
What are the colors of the Omani flag?

A: ، 1970 ديسمبر 17 الموافق هـ 1391 شوال 18 في مرة لأول ورفع سلطاني بقرار انشئ عمان لسلطنة الوطني العلم
. 1970 dysmbr 17 AlmwAfq h_ 1391 $wAl 18 fy mrp lA#wl wrfE slTAny bqrAr An$y# EmAn lslTnp AlwTny AlElm
The national flag of the Sultanate of Oman was established by a royal decree and was raised for the first time on Shawwal 18, 1391 AH corresponding to December 17, 1970.

Challenge: Diacritics are used in the question to note *which* Oman the speaker means; it is obvious from context in the article.

Vowel Diacritization in Arabic

Q: ؟ العُمانى العلم الوان هي ما
? AlEumAny AlElm AlwAn hy mA
What are the colors of the Omani flag?

A: ، 1970 ديسمبر 17 الموافق هـ 1391 شوال 18 في مرة لأول ورفع سلطاني بقرار انشئ عمان لسلطنة الوطني العلم
. 1970 dysmbr 17 AlmwAfq h_ 1391 $wAl 18 fy mrp lA#wl wrfE slTAny bqrAr An$y# EmAn lslTnp AlwTny AlElm
The national flag of the Sultanate of Oman was established by a royal decree and was raised for the first time on Shawwal 18, 1391 AH corresponding to December 17, 1970.

Challenge: Diacritics are used in the question to note *which* Oman the speaker means; it is obvious from context in the article.

Accents in Russian

Q:	Что	такое	атом	?			
	Chto	takoe	atom	?			
	What	such	atom	?			
	What is an		*atom*	*?*			
A:	Á том	—	частица	вещества	микроскопических	размеров	...
	Á tom	—	chastitsa	veschestva	mikroskopicheskih	razmerov	...
	Atom	PRED	particle	matter	microscopic	sizes	...
	An		*atom*	*is a microscopic particle of matter...*			

Challenge: Russian encyclopedia entries often use an accent to show emphasis. The question does not.

Morphological Inflection in Russian

Q:	Как	далеко	Уран	от	Земли	?
	Kak	daleko	Uran	ot	Zemli	?
	how	far	Uranus-SG.NOM	from	Earth-SG.GEN	?
	How far is		*Uranus*	*from*	*Earth*	*?*

A:	Расстояние	между	Ураном	и	Землёй	меняется	от	2,6	до	3,15	млрд	км…
	Rasstonyanie	mezhdu	Uranom	i	Zemlei	menyaetsya	ot	2,6	do	3,15	mlrd	km…
	distance	between	Uranus-SG.INSTR	and	Earth-SG.INSTR	varies	from	2,6	to	3,15	bln	km…
	The distance between		*Uranus*	*and*	*Earth*	*fluctuates from 2.6 to 3.15 bln km…*							

Challenge: Russian inflects main entities of interest differently in question versus answer.

Distinguishing Features

Dataset design

- Information *seeking:* Questions written without seeing the answer.
- No translation (no *Translationese*), no model-in-the-middle.
- Natural no-answer examples.
- Languages chosen to be typologically diverse (represent top ~100).

Intrinsic measures

- Lexical overlap is 3X lower vs XQuAD.
- Answer context 10X longer (article vs paragraph): difficult but useful.

Extrinsic measures

- F1 is far below human performance.
- 300k English SQuAD train worse than 10k English TyDi train.

Getting started - Experimenting is easy!

Blog post

Easy reading introduction.

(link)

Website

More glossed examples.

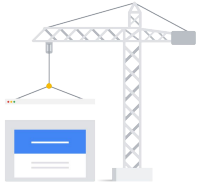
Leaderboard.

(link)

GitHub

Download data, train system, and evaluate... in **One Command**.

(link)



Code Guide

Code annotated to show you where to insert your new great idea!

(link)

Questions?

Thank you!