Crowdsourcing Question-Answer Meaning Representations

Julian Michael,[1] Gabriel Stanovsky,[1,3] Luheng He,[1] Ido Dagan,[2] and Luke Zettlemoyer[1]

github.com/uwnlp/qamr

[1]University of Washington

[2]Bar-Ilan University

[3]Allen Institute for Artificial Intelligence

Summary

Problem: semantic annotation requires experts, doesn't scale

Insight: Non-expert annotators can *use natural language to annotate natural language*

Challenge: get non-experts to provide high coverage of semantic relations in a sentence

Solution: A novel representation (QAMR), crowdsourcing setup, and dataset (>100k QA pairs)

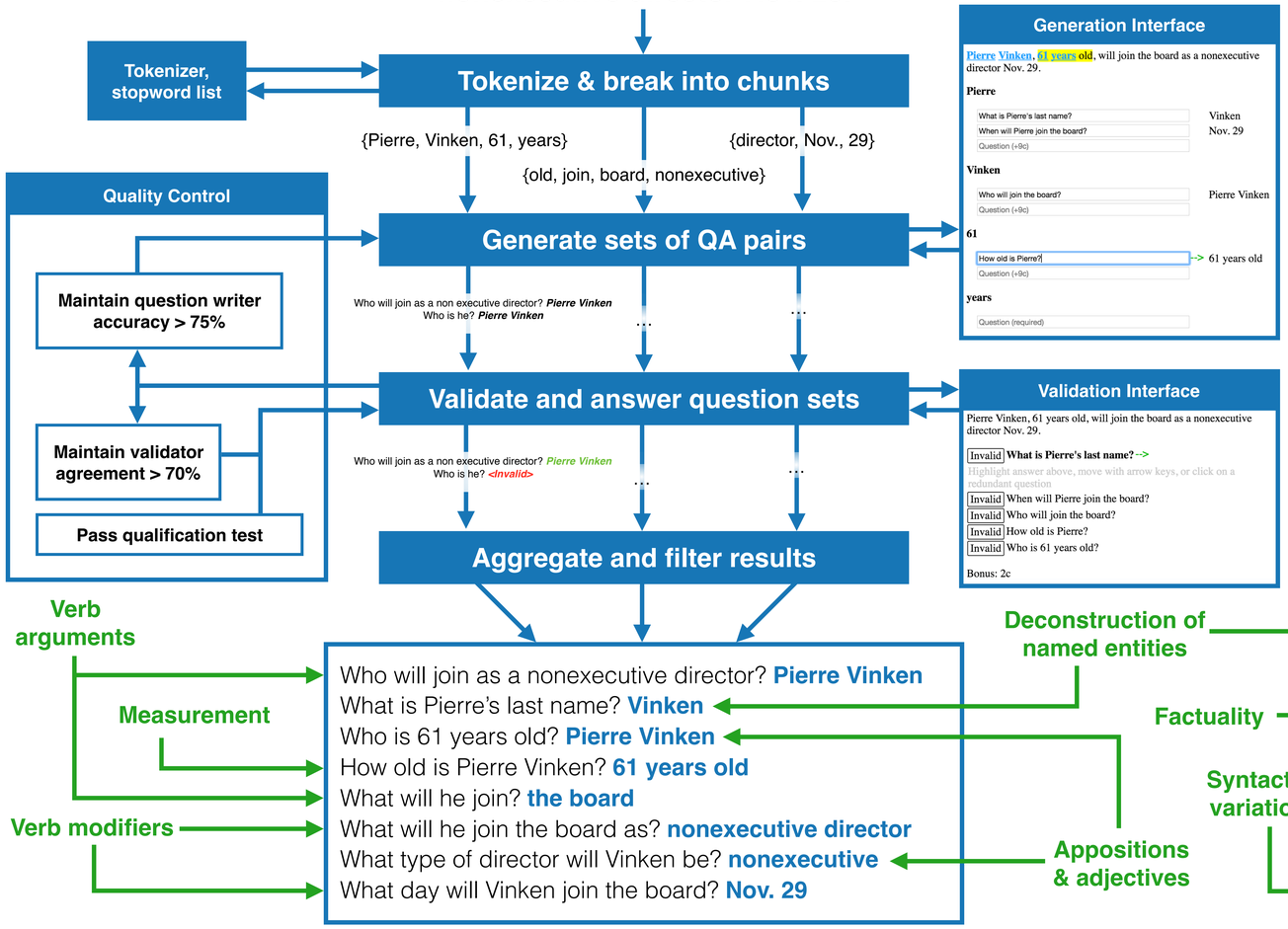
Related: differs from QA-SRL with *free-form* questions about *all predicate-argument relations*

Results: covers a wide range of semantic phenomena and can aid prediction of semantics in some settings (Open IE)

Follow-up: look for *Large-Scale QA-SRL Parsing* @ ACL!

Methods

Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.



Examples

Baruch ben Neriah, Jeremiah's scribe, used this alphabet to create the later scripts of the Old Testament.

Who wrote the scripts?
Baruch ben Neriah
Who did Baruch work for?
Jeremiah
What is old?
the Old Testament

← Open-vocabulary role labels

The ossicles are the malleus (hammer), incus (anvil), and the stapes (stirrup).

What is the malleus one of?
The ossicles

← Coreference

The situation is far from over as Sterling is refusing to sell the team, and the other teams have lobbied against him to force him to sell.

What is far from over?
The situation
Who are they lobbying against?
Sterling / him
Who wants to force him to sell?
other teams

← Semantic inference

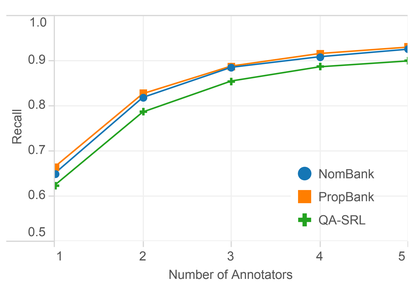
Mahlunga has said he did nothing wrong and Judge Horn said he "failed to express genuine remorse."

What is the Judge's last name?
Horn
Who doubted his remorse was genuine?
Judge Horn
Who didn't express genuine remorse?
Mahlunga

Climate change affects distribution of weeds, pests, and diseases.

What affects distribution of diseases?
Climate change
What is affected?
distribution of weeds, pests, and diseases

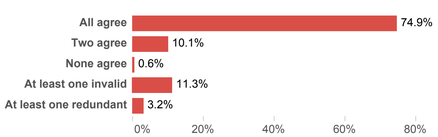
Data



Good question quality & answer agreement

High representational capacity: QAMR covers over 90% of predicate-argument relations in existing expert-annotated datasets, with sufficient annotation density

Coverage of predicate-argument relations remains relatively low in the training set, where we had 1 annotator for each target word



	PTB	Train	Dev	Test
Sentences	253	3938	499	480
Annotators	5	1	3	3
QA Pairs	18,789	51,063	19,069	18,959
Cost	$2,862	$7,879	$2,919	$2,919
Cost / Token	$0.44	$0.08	$0.25	$0.25

Low per-question annotation cost shows promise for scalable annotation of semantic structure with crowdsourcing

Modeling

Question Generation

Choose seed (w_i, w_j)

Encode sentence with seed

Decode question with copying

.28 P / .24 R fuzzy match

Learning to generate diverse sets of free-form questions is challenging

Question Answering

Pool training set with SQuAD

Out-of-the-box reading comprehension model (BiDAF)

75.7% EM / 83.9% F1

QA is relatively easy: most of QAMR's signal is in the questions

Open IE*

Convert QAMR to Open IE & pool with existing data

BiLSTM sequence tagger

SoTA Open IE system

*Concurrent work: *Supervised Open Information Extraction*, Stanovsky et al., NAACL 2018

QAMR is successful when parsed and applied appropriately