BERT + Verify

Kai Marshland
kaimarsh@stanford.edu

Problem

- Answering questions!
- SQuAD 2.0 dataset
- Sometimes there isn't an answer

But wait! Answering questions is a different task than figuring out whether or not the question has an answer in the first place, yet most current approaches treat it as the same thing.

Approach

Add a second verification step after the answer generation step. If it thinks the question is unanswerable, it will return no-answer no matter what the generator thinks.

Verifier #1: Generative Pre-trained Transformer
Verifier #2: BERT

Question & answer get fed through these complex models, then through a linear layer to be classified.

Analysis

Predicts no-answer too often. The algorithm only really increases the amount of no-answers, which overall hurts performance.

The BERT-based verifier is *slow*. An order of magnitude slower than without a verifier.

It adds too many parameters that take up space, make it slow, and don't get us much.

Data

Questions + Context

"The flora of the city may be considered very rich in species. The species richness is mainly due to the location of Warsaw within the border region of several big floral regions comprising substantial proportions of close-to-wilderness areas (natural forests, wetlands along the Vistula) as well as arable land, meadows and forests. Bielany Forest, located within the borders of Warsaw, is the remaining part of the Masovian Primeval Forest. Bielany Forest nature reserve is connected with Kampinos Forest. It is home to rich fauna and flora. Within the forest there are three cycling and walking trails. Other big forest area is Kabaty Forest by the southern city border. Warsaw has also two botanic gardens: by the Łazienki park (a didactic-research unit of the University of Warsaw) as well as by the Park of Culture and Rest in Powsin (a unit of the Polish Academy of Science).

Long and boring -- we want to automate it

Why is Warsaw's flora very rich in species ?

Results

Model	F1 Score	EM Score
Baseline (30 epochs)	60.49	57.26
GPT verifier (30 epochs)	52.19	52.19

GPT-based verifier is *bad*

Model	F1 Score	EM Score
Baseline (3 epochs)	51.89	49.04
BERT verifier (3 epochs)	49.94	49.68

BERT-based verifier is *okay*

Conclusions

Verifiers aren't worth it

- They're slow
- Gains are small
- Answer generation is already really powerful

The core assumption that a verifier depends upon seems wrong. Though it might make intuitive sense to handle the problems of answer generation and answerability separately, they both depend on similar types of understanding, and should not be separated

References

[1] M. Hu, F. Wei, Y. Peng, Z. Huang, N. Yang, and D. Li, "Read + verify: Machine reading comprehension with unanswerable questions," 2018.
[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.
[3] F. Sun, L. Li, X. Qiu, and Y. Liu, "U-net: Machine reading comprehension with unanswerable questions," 2018.
[4] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," 2016.