Controlling Information Aggregation for Complex Question Answering

Heeyoung Kwon[1], Harsh Trivedi[1], Peter Jansen[2], Mihai Surdeanu[2]
, and Niranjan Balasubramanian[1]

[1] Stony Brook University, Stony Brook NY 11790, USA
[2] University of Arizona, Tucson AZ 85721, USA
{heekwon, hjtrivedi, niranjan}@cs.stonybrook.edu,
{pajansen, msurdeanu}@email.arizona.edu

Abstract. Complex question answering, the task of answering complex natural language questions that rely on inference, requires the aggregation of information from multiple sources. Automatic aggregation often fails because it combines semantically unrelated facts leading to bad inferences. This paper proposes methods to address this inference drift problem. In particular, the paper develops unsupervised and supervised mechanisms to control random walks on Open Information Extraction (OIE) knowledge graphs. Empirical evaluation on an elementary science exam benchmark shows that the proposed methods enables effective aggregation even over larger graphs and demonstrates the complementary value of information aggregation for answering complex questions.

1 Introduction

Question answering (QA), i.e., finding short answers to natural language questions, is one of the most important but challenging tasks on the road towards natural language understanding [Etzioni, 2011]. QA methods are moving beyond tackling simple factoid questions to more complex questions, which require aggregating and reasoning over multiple pieces of information. The elementary science exam benchmark, for example, includes questions that test the student's ability to reason over connected facts [Clark et al., 2013, Clark and Etzioni, 2016, Jansen et al., 2016]. Such aggregation however is prone to inference drift, where semantically unrelated pieces of information are combined resulting in spurious inferences. This is especially true when reasoning over knowledge graphs built using shallow semantic representations as Open Information Extraction (OIE) triples [Mausam et al., 2012]. To mitigate this issue, recent QA methods that operate over OIE graphs limit themselves to reasoning with artificially short paths [Khot et al., 2017, Khashabi et al., 2017].

Rather than limit inference to smaller paths or sub-graphs, this paper proposes methods that allow for reasoning over larger graphs while still avoiding inference drifts. To this end, this paper formulates complex QA as a random walk method over knowledge graphs formed with OIE triples. Each random walk begins at some question nodes, and aggregates information by following relation edges to reach other nodes. For example, for a question asking if an iron nail is a conductor or an insulator, a random walk explaining the positive answer to this question begins with iron nail, and links facts such as `(iron nail, is made of, iron)`, `(iron, is a, metal)`, and `(metals, are, electric conductors)`, all obtained from different sentences. Answer nodes are then ranked by the random walk scores computed over these

paths. It is easy to imagine how this type of linking can lead to inference drift, where semantically irrelevant pieces of information for QA are combined so that causing spurious inferences, especially on large graphs.

Our main idea to minimize inference drift is to actively guide the random walks to stay on paths that are relevant to the question context. We explore two instantiations of this idea with PageRank [Haveliwala, 2002]: (i) We develop unsupervised estimations of node importance, edge and teleportation probabilities to guide the random walks. These are estimated based on some measure of similarity of the nodes and edges with respect to the overall question context. (ii) We develop a novel supervised PageRank formulation that directly estimates the node importance, edge and teleportation probabilities. We use a set of retrieval features to model node and edge importance and use supervision to combine these in order to directly maximize the end QA performance.

Our empirical evaluations on a standard science exam benchmark shows that: (i) controlled aggregation with drift-sensitive PageRank yields up to +3.2% gain in accuracy (precision@1) over standard topic-sensitive PageRank (TPR) (ii) the new supervised formulation yields even better results with a +3.7% gain over TPR, (iii) aggregation over sentences with drift-sensitive methods improves over a sentence-only model in an ensemble with a +2.0% gain. Overall, the results show that aggregation can be useful for complex questions when it is controlled carefully to stay in the question context.

2 Graph-based Reasoning for QA

Complex QA can be formulated as a reasoning problem over knowledge graphs that aggregate related facts. Given an appropriate graph, finding an answer translates into a traversal of relevant facts that lead to the answer. In the case of multiple-choice questions, this can be cast as a ranking problem among nodes corresponding to the candidate answer choices. The idea is to induce a knowledge sub-graph that connects the question related nodes to candidate answer choices and then assess the strength of these connections. The implicit assumption is that a correct answer has stronger connections to the question nodes than incorrect answers. Figure 1 illustrates with an example question from a science exam. Nodes in this graph are question terms (blue), answer choices (green for correct and pink for incorrect), and other concepts that link the question and answer choices through OIE relations. As it shows, with automatically extracted OIE relations, the graph can include nodes and relations that cause inference drift (shown via the red node and dashed arrows). For example, even when the random walk begins from a question node, say "walking", inference can erroneously link `(central location, is great for, walking)`, `(heart, is a, central location)`, and `(heart, is a type of, muscle)` to conclude that `(heart, is used for, walking)`.

We adapt Topic-sensitive PageRank (TPR) to assess the importance of the answer nodes. Consider a walker who starts from one of the question nodes and follows the graph structure. At any node i, the walker can either choose one of the outgoing edges with probability $(1 - d_i)$ or ignore the outgoing edges and instead teleport to one of the question nodes (i.e., the seed nodes) with probability d_i. The seed probability v_j specifies the likelihood that the random walker would choose to teleport to the node j (we discuss later multiple strategies for initializing seed probabilities). Intuitively, we

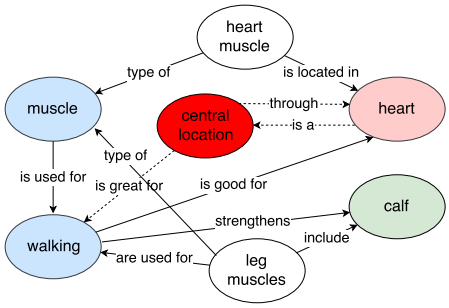

Fig. 1. OIE based knowledge graph for the question: `Which muscle is used for walking? (A) heart (B) calf`

use the seed probabilities to bias the random walks towards question nodes. This ensures that nodes well connected to question nodes will get higher PageRank (PR) scores.

Given the transition matrix A, the seed probabilities vector v, and the teleportation probabilities vector d, the PR vector π at time step $t+1$ is: $\pi^{(t+1)} = (1-d)A\pi^t + dv$. The answer choice that receives the highest PR score is returned as the answer.

2.1 Drift-Sensitive Page Rank

Even with TPR walks that are already biased towards question nodes, inference drift remains an issue. Effective solutions should ensure that reasoning stays close to paths that are relevant to specific question context. We propose two such solutions here.

Unsupervised Estimation Our first method introduces node-specific teleportation probabilities that enhance the likelihood that teleportation actions land on nodes related to the question context. The method operates as follows. (i) We estimate *transition probabilities* by scoring the edges based on the relevance score given to the source sentence by the information retrieval model. These scores are then normalized by their sum to turn them into probabilities. (ii) We estimate *node-specific teleportation* based on their importance within the question context. That is, when a walk reaches a high-relevance node, we want it to continue. When it reaches a low relevance node, however, we want it to teleport and restart from one of the question nodes. Thus we want the teleportation probability to be low when the relevance is high and vice versa. We experiment with the degree of the node and the similarity of the node to the question context as surrogates to estimate relevance (see Section 3). (iii) We use *seed probabilities* to capture relative node importance. Specifically, we use normalized scores from the abstractness-concreteness norms, a measure that was previously shown to be useful for measuring whether a word in the question is likely to be a key focus word [Jansen et al., 2017].

Supervised Page Rank Our second method learns a parameterization of TPR that improves the ranking of correct answers. PageRank's effectiveness depends on the quality of the graph structure and the probability estimates used in the random walk: the transition (A), seed (v), and teleportation (d) probabilities. Rather than using independent estimation methods, we develop a supervised PR method that learns a non-linear function, via a two layer feed forward network, which combines well-known information retrieval features to compute the estimates. This supervised version can be written as a parametrization[3] of TPR as shown as follows: $\pi^{(t+1)} = (1-d)A_\phi\pi^t + dv_\theta$, where

[3] This is inspired by the work of Gao et al. [2011], who use a linear parametrization but for a single graph problem using a different modeling approach.

θ is the parameter vector for seed probability features and ϕ the parameter vector for edge probability features. For seed probability estimation, we use three features: (1) focus word weights [Jansen et al., 2017], (2) Rocchio query expansion weights [Rocchio, 1971], a traditional measure of relative term importance, and (3) an entropy based discriminativeness measure, that ensures a question word that occurs in all answer choices receives a low score, and one that occurs in only one answer choice gets a high score. For edge features, we use (1) IR score, which is retrieval score of triple from IR system when queried with question text, (2,3) Context Score of source and target (word-wise word2vec similarity between question and each node with entailment score for out-of-vocabulary), (3,4) Entropy based discriminativeness score of source and target, (5) Triple confidence score from OIE extraction, (6-9) boolean features for edge type (QI, IA, AI, II)[4].

Let x_s denote the seed probability features for a node s, and let z_{st} denote the edge probability features for an edge (s, t). Then, we have: $v_\theta(u) = \frac{f_\theta(x_s)}{\sum_{u \in G} f_\theta(x_u)}$ and $A_\phi(s,t) = \frac{g_\phi(z_{st})}{\sum_{e_{sq} \in G} g_\phi(z_{sq})}$ The functions $f_\theta(.)$ and $g_\phi(.)$ are feed-forward networks with 1 hidden layer of 3 nodes. Output layer has 1 node and Rectified Linear Unit (ReLU) is used for the activation units. Weights θ and ϕ control node specific reset probabilities and edge specific transition probabilities in neural networks f and g respectively. Training finds parameters that maximize the following function:

$$\arg\max_{\theta,\phi} \sum_{q} \left((log(\pi(a_c)) - log\big(\sum_{a_i \in ans(q)} \big(\pi(a_i)\big) \right)$$

This objective cannot be optimized in closed form but since it is differentiable we use Adam, a stochastic optimization method [Kingma and Ba, 2014].

In contrast to the unsupervised setting, the bidirectional edges were unhelpful in the supervised setting due to stability issues in training. We report results on the graph where we remove backward edges (IQ, AQ and AI) for improved convergence.

3 Evaluation

We evaluate our approach for aggregation on a standard science exam benchmark[5], which consists of multiple choice questions for 6th-9th grade science. The dataset set includes 2,068 questions for training, of which we reserve 485 as a development set (dev) and a blind test set of 1,639 questions. We build knowledge graphs using OIE triples from five relevant corpora including study guides, an openly available textbook, and study flashcards totaling 588,472 triples.

3.1 Drift-sensitive PageRank

Table 1 shows the accuracy of the drift-sensitive PageRank methods. All models were run on graphs built over top K sentences returned by a sentence retrieval model [Jansen et al., 2017]. We set K = 40 based on dev set performance (see Table 2 for other K values). In order not to end up with several fragmented graphs, we introduced entailment edges. We only kept the high scored edges and the threshold is tuned also based on dev set performance. All drift-sensitive methods perform better than regular PR (top

[4] QI denotes edge between Question and Intermediate Node

[5] https://www.kaggle.com/c/the-allen-aiscience-challenge

row) and TPR (second row). Using question words as seeds with uniform probabilities provides a +3.17 gain in accuracy over TPR. Using focus word weights as seed probabilities yields an additional +1.56 points. Using teleportation estimates based on question context (lower chance of teleportation when node is more similar to question) gives a +2.8 points gain. Supervised PageRank outperforms all variants with a +3.66 gain over basic TPR and +0.85 gain over the best unsupervised method.

method	seeds	teleportation	test	reference
page rank	none	uniform	35.51	
TPR	uniform	uniform	38.26	
	focus	uniform	40.33	(A)
drift-sensitive	focus	quest. sim.	41.49	(B)
	sup.	sup.	42.34	Sup.

Table 1. Comparison of different drift-sensitive methods: focus - Using abstract/concreteness norms based probabilities. quest. sim. - Using question similarity for teleportation probabilities

3.2 Graph Sizes

We evaluated performance with knowledge graphs of different sizes by varying the number of source sentences (Table 2). Drift-sensitive methods perform better than the TPR across all data sizes, with more pronounced differences when the graph sizes are larger. Inference drift is more likely in large graphs and controlling random walks with drift-sensitive methods helps in these cases.

method	top 10	top 20	top 30	top 40	top 50
TPR	39.54	40.63	41.31	38.26	38.68
unsupervised	41.00	41.55	42.46	40.33	39.84
supervised	41.30	42.22	41.80	42.34	42.40

Table 2. Performance on different graph sizes using top X sentences to construct the graph

3.3 Utility of Aggregation

Aggregation is not only useful on its own, but it provides complementary benefits in an ensemble with non-aggregation based methods. We built an ensemble model with a strong non-aggregation model, a supervised sentence-retrieval model trained in [Jansen et al., 2017]. The ensemble simply chooses between the two models based on their scores and the number of nodes and edges in the concept graphs. Most drift-sensitive models add value in this simple ensemble, with the supervised model providing a +2 points gain over the sentence model thus showing the complementary value of aggregation for complex questions (Table 3).

method	sent	sent + (A)	sent + (B)	sent + Sup.
accuracy	43.44	44.30	45.58	45.45

Table 3. Results of aggregation with a sentence retrieval model (sent)

4 Related Work

For the elementary science benchmarks with complex questions, a range of inference methods have been explored, including probabilistic reasoning with first-order logic [Khot et al., 2015], constraints-based inference with semantic and shallow semantic structures [Clark et al., 2016, Khot et al., 2017], and graph-based methods with syntactic alignment and lexical semantics [Sharp et al., 2015]. Sharp et al. [2015] show a method

for aggregating information from multiple sentences using syntactic structure based alignments. Khot et al. [2017] show how OIE can benefit a constraint-based inference mechanism that tightly controls how multiple short facts can be combined. Fried et al. [2015] show that graphs built using words or syntactic dependencies can aggregate knowledge to improve performance on a community question answering task, but that long graph traversals lead to "semantic drift" and decreased performance. Our work builds on these ideas but explores the utility of the different graph methods explicitly through the use of page rank based methods.

5 Conclusions

Aggregating information from multiple texts is critical for answering complex questions. However, aggregation introduces spurious inferences, especially when using shallow semantic representations such as Open Information Extraction graphs. This forces models to limit reasoning to smaller paths or sub-graphs. Instead this paper introduces drift-sensitive variants of PageRank that allow for effective reasoning over large graphs. By controlling the random walks to stay on question contexts, the drift-sensitive methods achieve substantial gains over standard topic-sensitive PageRank and provides gains in an ensemble with a sentence-level model.

References

O. Etzioni. Search needs a shake-up. *Nature*, 476(7358):25–26, 2011.

Peter Clark, Philip Harrison, and Niranjan Balasubramanian. A study of the knowledge base requirements for passing an elementary science test. In *AKBC@CIKM*, 2013.

Peter Clark and Oren Etzioni. My computer is an honor student - but how intelligent is it? standardized tests as a measure of ai. *AI Magazine*, 37:5–12, 2016.

Peter Jansen, Niranjan Balasubramanian, Mihai Surdeanu, and Peter Clark. What's in an explanation? characterizing knowledge and inference requirements for elementary science exams. In *COLING*, 2016.

Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. Open language learning for information extraction. In *EMNLP-CoNLL*, 2012.

Tushar Khot, Ashish Sabharwal, and Peter Clark. Answering complex questions using open information extraction. *CoRR*, abs/1704.05572, 2017.

Daniel Khashabi, Tushar Khot Ashish Sabharwal, and Dan Roth. Learning what is essential in questions. *COLING*, 2017.

Taher H. Haveliwala. Topic-sensitive pagerank. In *WWW*, 2002.

Peter Jansen, Rebecca Sharp, Mihai Surdeanu, and Peter Clark. Framing qa as building and ranking intersentence answer justifications. In *Computational Linguistics*, 2017.

Bin Gao, Tie-Yan Liu, Wei Wei, Taifeng Wang, and Hang Li. Semi-supervised ranking on very large graphs with rich metadata. In *ACM SIGKDD*, pages 96–104, 2011.

J. J. Rocchio. Relevance feedback in information retrieval. *Salton: The SMART Retrieval System: Experiments in Automatic Document Processing*, 1971.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Tushar Khot, Niranjan Balasubramanian, Eric Gribkoff, Ashish Sabharwal, Peter Clark, and Oren Etzioni. Exploring markov logic networks for question answering. In *EMNLP*, 2015.

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter D. Turney, and Daniel Khashabi. Combining retrieval, statistics, and inference to answer elementary science questions. In *AAAI*, 2016.

Rebecca Sharp, Peter Jansen, Mihai Surdeanu, and Peter Clark. Spinning straw into gold: Using free text to train monolingual alignment models for non-factoid question answering. In *HLT-NAACL*, 2015.

Daniel Fried, Peter Jansen, Gustave Hahn-Powell, Mihai Surdeanu, and Peter Clark. Higher-order lexical semantic models for non-factoid answer reranking. *Transactions of the Association for Computational Linguistics*, 3:197–210, 2015.