Learning Knowledge Graphs for Question Answering through Conversational Dialog

Ben Hixon
University of Washington
Computer Science & Engineering
Seattle, WA, USA
bhixon@cs.washington.edu

Peter Clark
Allen Institute for AI
Seattle, WA, USA
peterc@allenai.org

Hannaneh Hajishirzi
University of Washington
Electrical Engineering
Seattle, WA, USA
hannaneh@washington.edu

Abstract

We describe how a question-answering system can learn about its domain from conversational dialogs. Our system learns to relate concepts in science questions to propositions in a fact corpus, stores new concepts and relations in a knowledge graph (KG), and uses the graph to solve questions. We are the first to acquire knowledge for question-answering from open, natural language dialogs without a fixed ontology or domain model that predetermines what users can say. Our relation-based strategies complete more successful dialogs than a query expansion baseline, our task-driven relations are more effective for solving science questions than relations from general knowledge sources, and our method is practical enough to generalize to other domains.

1 Introduction

Our goal is a question-answering (QA) system that learns about its domain from users in open, natural language dialogs. Dialog learning interfaces built on top of existing QA systems can complement and refine knowledge acquired from other sources. Even large QA systems such as Watson (Ferrucci et al., 2010) that utilize heavy domain engineering could benefit from focused dialogs to dynamically improve their knowledge.

A successful QA system requires domain knowledge. For example, 4th grade science questions are difficult since they often exclude knowledge necessary to relate answers to known facts. The question in Figure 1 asks if an *iron nail* conducts electricity;

S1: Here's the question: Which object is the best conductor of electricity? (A) a wax crayon (B) a plastic spoon (C) a rubber eraser (D) an iron nail. I'm not sure of the answer. What do you think the answer is?
U1: iron nail
S2: Why do you think the answer is (D) an iron nail?
U2: electricity goes through metal
S3: I know "metals are shiny and good conductors of electricity and heat," does that mean that the answer is (D) an iron nail?
U3: yeah
S4: Okay, great. What's the relationship between 'electricity' and 'iron', if any?
U4: iron conducts electricity because its metal
S5: Okay, great, I get it! The answer is (D) an iron nail

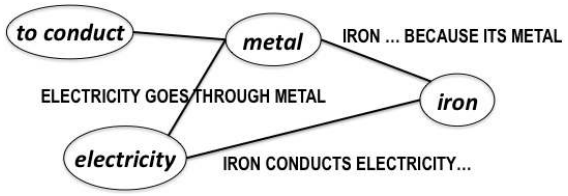


Figure 1: Top: A successful real user dialog. Open-ended prompts (S1&S2) encourage conversational explanations. Focused prompts (S4) target particular relations. Bottom: Corresponding knowledge graph consisting of relations between concepts.

the system only knows that *metal* conducts electricity, and it needs to learn that *iron* is a *metal* in order to answer the question with the relevant fact.

Our dialog system, KNOWBOT, conducts dialogs about science questions and learns how concepts in each question relate to propositions in a corpus of science facts. KNOWBOT presents its user with a question (line S1 in Figure 1), prompts them to choose and explain their answer, and extracts *relations* – any semantic relationship between two con-

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 851–861,
Denver, Colorado, May 31 – June 5, 2015.

cepts, such as *metal* to *iron* (line U4 in Figure 1) – that increase its confidence in the user's answer.

Relation extraction systems such as NELL (Carlson et al., 2010) use ontologies to predetermine valid relation types and arguments, then scan text to fill the ontology with facts. Open Information Extraction (Etzioni et al., 2011) avoids fixed ontologies with domain-independent linguistic features, distant supervision, and redundancy, but requires web-scale text and doesn't improve with interaction. Like Open IE, we extract relations without predetermined types, but are the first to do so from dialog.

KNOWBOT is an *open* dialog system, which means a user utterance may progress the dialog task even if its underlying action is not explicitly represented in a dialog model. This lets KNOWBOT quickly bootstrap domain knowledge from users without significant engineering overhead. Dialog-driven extraction produces effective relations without annotation, improves after each interaction, acquires relations useful on a particular task, and embeds relations in a rich dialog context.

Users successfully correct the system in approximately 50% of dialogs even without a predetermined dialog model. A baseline query expansion (Bast et al., 2007) strategy that bases decisions on the acquisition of new keywords instead of new relations results in only a 5% success rate. In comparison to paraphrase relations from general knowledge bases, relations acquired by our method are more effective as domain knowledge, demonstrating that we successfully learn from real users.

Our contributions include:

1. The first end-to-end system to construct knowledge graphs for question-answering through conversational dialog.

2. A generalizable method to represent the meaning of user utterances without a dialog model when task progression can be computed as a function of extracted relations.

3. A novel data set of real user dialogs in which users correct a QA system's answer, together with knowledge graphs representing the important concepts and relations in each question, labeled with rich dialog features.

2 Conversational extraction for QA

Our QA task consists of 107 science questions from the 4th grade New York Regents exam (Clark et al., 2014).[1] Each question has four possible answers. We convert each of the four question-answer pairs into a true/false *question-answer statement* using a small number of pattern-based transformation rules.

Just as 4th graders read their textbooks for answers, we collect SCITEXT (Clark et al., 2014), a corpus of unlabeled true-false natural language sentences from science textbooks, study guides, and Wikipedia Science. Each question-answer statement is associated with a subset of true/false support sentences from SCITEXT based on positive word overlap between the question-answer pair and the support sentence. The degree to which a SCITEXT sentence supports a question-answer pair is the sentence's *alignment score* (section 2.3).

Initially, the alignment score depends on keyword overlap alone, but SCITEXT needs domain knowledge to answer our questions. For example, the correct question-answer statement to *What form of energy causes an ice cube to melt? (A) mechanical (B) magnetic (C) sound (D) heat* is $Q_{(D)}$, "Heat is a form of energy and heat causes an ice cube to melt." To better align $Q_{(D)}$ to the SCITEXT sentence "A snowball melting in your hand is an example of heat energy," we need to know that *snowballs* are made of *ice*. Figure 2 illustrates this example.

To construct a knowledge base with which to use SCITEXT, we extract *concepts* (section 2.1) from questions and SCITEXT sentences, then use *relations* (section 2.2) between concepts to determine which question-answer statement Q_i is most highly aligned with a supporting SCITEXT sentence.

2.1 Concepts

A *concept keyword* in a sentence or user utterance is any non-stopword of at least three characters. Stopwords are domain-independent, low-information words such as "the."

A *concept* is a set of concept keywords with a common root, e.g. {*melts, melted, melting*} or {*heat, heats, heated*}. We use the Porter algorithm for stemming (Porter, 1997). *Question concepts* ap-

[1] Our dialogs, extractions, and tools are available at `www.cs.washington.edu/research/nlp/knowbot`

pear in a question-answer statement, and *support concepts* appear in a SCITEXT support sentence.

2.2 Relations

A *relation* is any pair of concepts that represents a semantic correspondence. In general, relations can be labeled with any feature that describes the correspondence, such as a particular *type*. For example, the relation between *Obama* and *Hawaii* can be labeled with the type *born-in*.

A predetermined ontology is typically required to label relations with their type. In this work we label acquired relations with dialog-specific features. Our thesis is that user explanations intend to relate concepts together, and the system's task is to determine the user's intent. For example, the user utterance `U: it's melting because of heat` relates the concepts represented by *melt[ing]* and *heat*, with the words `because of` appearing between the two concept keywords. We refer to `because of` as the relation's *intext*.

Relations can be intuitively arranged as a *knowledge graph*, which in this work is any graph whose nodes are concepts and whose edges are relations between those concepts, in the spirit of semantic networks such as ConceptNet (Havasi et al., 2007).

2.3 Sentence alignment

We calculate the alignment score α between the ith question-answer statement Q_i and its jth supporting SCITEXT sentence $S_{i,j}$ as the normalized number of relations between their concepts,

$$\alpha(Q_i, S_{i,j}) = \frac{\|R_{Q_i,S_{i,j}}\|}{\|C_{Q_i} \cup C_{S_{i,j}}\|}, \qquad (1)$$

where C_{Q_i} is the set of concepts in Q_i, $C_{S_{i,j}}$ is the set of concepts in $S_{i,j}$, and $\|R_{Q_i,S_{i,j}}\|$ is the number of relations between C_{Q_i} and $C_{S_{i,j}}$.

Normalized relation count is a practical semantic similarity score that generalizes to different knowledge representations. The dialog in Figure 2 aligns $Q_{(D)}$ with the SCITEXT fact S by learning from the user that, for example, *heat* is related to *melting*.

3 The KNOWBOT dialog system

KNOWBOT grows a knowledge graph of commonsense semantic relations in open, conversational dialog. Figure 2 traces the growth of a knowledge graph over a single dialog. Section 3.1 details how knowledge is extracted from user explanations without a dialog model. Section 3.2 describes dialog strategies that elicit natural language explanations.

KNOWBOT uses task progress to drive natural language understanding. It assumes the user intends to provide one or more novel relations, and uses the constraints described in section 3.1.1 to disambiguate noisy relations. This way, KNOWBOT knows when the dialog progresses because its confidence in the user's chosen answer increases.

3.1 Building knowledge graphs from dialog

KNOWBOT builds KGs at three levels: per utterance, per dialog, and globally over all dialogs. An *utterance-level knowledge graph* (uKG) (Figure 2a) is a fully connected graph whose nodes are all concepts in an utterance. After aggressive pruning (section 3.1.1), remaining edges update a *dialog-level knowledge graph* (dKG) (Figure 2b; section 3.1.2).

Upon dialog termination, the dKG updates the *global knowledge graph* (gKG), which stores relations acquired from all dialogs (section 3.1.3).

3.1.1 Utterance-level KGs

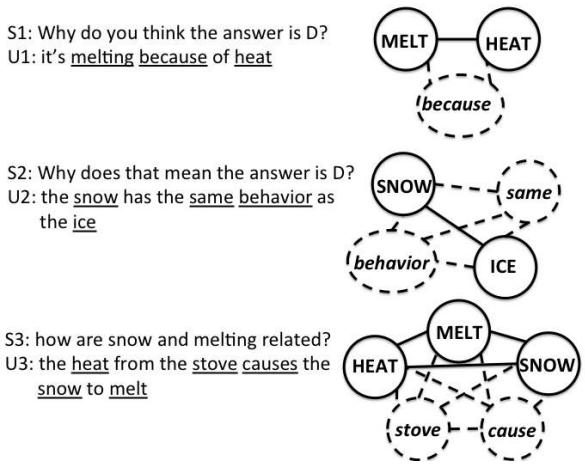
KNOWBOT initially relates every pair of concepts in an utterance, then prunes them based on two constraints: *alignment* and *adjacency*.

Each user explanation is first converted into a fully-connected utterance-level knowledge graph. This uKG is noisy because users don't intend relations between every pair of keywords in their utterance. For example, a typical utterance `U: freezes means it changes water from a liquid to a solid` mentions six concepts, *freezing, meaning, change, water, liquid, solid*, with $\binom{6}{2}$ potential binary relations. Not every relation is salient to the question. To remove noisy relations, edges in the uKG are aggressively pruned with two simple, rational constraints:

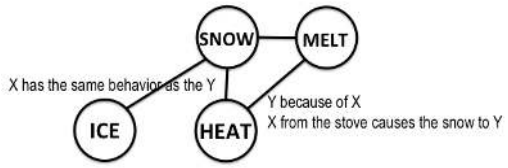
1. *Alignment*. An edge can only relate a question concept to a support concept.

2. *Adjacency*. Edges can't relate concepts whose keywords are adjacent in the utterance.

The intuition for the alignment constraint is that the user intends each explanation to relate the question



(a) utterance-level knowledge graph (uKG)



(b) dialog-level knowledge graph (dKG)

Q = heat is a form of energy and heat causes an ice cube to melt
S = a snowball melting in your hand is an example of heat energy

(c) The dialog goal is to align Q and S

Figure 2: Every pair of concepts in each user utterance is related then aggressively pruned. (a) Utterance-level knowledge graphs represent individual utterances. Concepts (underlined, inset in nodes) are obtained by removing stopwords and stemming. An edge that either doesn't connect a question and support concept or else which connects concepts whose keywords in the user utterance have no intervening words (*intexts*) are pruned, indicated here with dashed lines. (b) The four remaining relations are stored in a dialog-level dKG.

to a known fact, and other relations in the utterance are unintentional. For example, in the uKG for the first utterance in Figure 2(a), the edge between *melt[ing]* and *heat* is an alignment relation because *melt[ing]* is a concept in S and *heat* is a concept in Q. But the edge between *because* and *heat* is pruned (dashed lines) since *because* is not a concept in S.

Adjacency is a simple, practical syntactic feature to reduce spurious relations. Users typically put words or *intexts* between concepts they intend to relate. The edge between *melt[ing]* and *because* is pruned since their keywords are adjacent in `U1: it's melting because of heat`, while `U2` relates *snow* and *ice* with the intext `has the same behavior as the`.

We find these constraints effective in practice, but at this point other pruning constraints can be deployed. A strength of our approach is that it welcomes aggressive pruning: just as in human-human interaction, users who initially fail to communicate their intention can try again later in the dialog.

3.1.2 Dialog-level KGs

Each dialog focuses on a single question. KNOWBOT starts with an empty dialog-level knowledge graph (dKG). After each user turn, edges from that turn's uKG are added to the dKG, and KNOWBOT rescores each of the four answers according to equation (1) where the set of relations $R_{Q_i,S_{i,j}}$ is exactly the set of edges in the dKG. The dialog successfully terminates when the user's answer has the highest alignment score, indicating the "missing knowledge" has been successfully provided by the user.

3.1.3 The global knowledge graph

The *global knowledge graph* (gKG) includes every relation learned from every KNOWBOT dialog.

Because we do not use a fixed ontology or comprehensive dialog model, individual dialogs can result in noisy relations even after aggressive pruning. However, as KNOWBOT conducts more dialogs about the same problem, relations that more often re-occur are more likely to be salient to the problem.

In this work, KNOWBOT takes advantage of redundancy with a simple filter: it ignores *singleton* relations originating in a single user utterance. We find even this simple filter increases performance. As KNOWBOT accumulates more dialogs, frequency can be incorporated in more sophisticated models.

3.2 Dialog strategies for knowledge acquisition

We've described how a user's free text explanations are converted into knowledge graphs. Each user explanation is uttered in response to a system

prompt. A dialog system's *dialog manager* chooses the prompt to say next according to its *dialog strategy*, which maps each system state to an action. An effective dialog strategy guides users to informative explanations that provide novel relations which let KNOWBOT successfully answer the question.

We compare two different strategies. A *user-initiative* strategy always asks open-ended questions to prompt the user for new explanations, e.g. line S2 in Figure 1. These prompts let users introduce salient concepts on their own.

In contrast, a *mixed-initiative* strategy utilizes focused prompts (line S4 in Figure 1) to introduce potentially related concepts. KNOWBOT chooses what pair of concepts to ask about based on how *discriminative* they are. The most discriminative concepts are the pair of question and support concepts that (1) don't already have an edge between them, (2) satisfies the alignment constraint for the user's answer, and (3) satisfies the alignment constraint for the fewest alternative answers. By proposing relations that would lead to a swift completion of the dialog task, KNOWBOT shares the burden of knowledge acquisition with the user.

Both dialog strategies are question-independent, but because we don't use a comprehensive dialog model to represent the state space, we rely on hand built rules instead of optimizing with respect to a reward function. For example, KNOWBOT always starts by asking the user for their answer, and if a new support sentence is found will always immediately present it to the user for confirmation.

4 Evaluation of dialog strategies

Our first experiment compares mixed-initiative and user-initiative strategies (section 3.2) to a baseline interactive query expansion (section 4.1). The purpose of this experiment is to investigate whether users can successfully complete our complex dialog task even though we don't use a trained semantic parser for natural language understanding.

Dialogs were conducted through a web browser. Users were colleagues and interns at the Allen Institute for Artificial Intelligence, and so were familiar with the question-answering task but were not expert annotators. Users were invited to converse with the system of their choice, and to move on to a new question if they felt the dialog was not progressing. Individual dialog sessions were anonymous.

The system starts each dialog with an empty knowledge graph, using only *identity* relations to select its answer. This default answer is correct on 44 of the 107 questions, and an additional 10 questions have no associated supporting sentence for the correct answer in SCITEXT. We run dialogs for the remaining 53 questions, for which each answer candidate has 80 supporting sentences in SCITEXT on average. A successful dialog terminates when the system extracts enough novel relations from the user that the correct answer has the highest alignment score with one of its supporting sentences.

4.1 Baseline: Interactive query expansion

To evaluate whether task-driven relation extraction is an effective method for knowledge acquisition in the absence of an explicit dialog model, we also implement a baseline dialog strategy based on interactive *query expansion* (IQE). This baseline is similar to the recent knowledge acquisition dialog system of Rudnicky and Pappu (2014a; 2014b).

In IQE, new knowledge is learned in the form of novel keywords that are appended to the question-answer statement. For example, the dialog in Figure 1 shows the user teaching KNOWBOT how *metal* relates to *electricity*. KNOWBOT understands that the user intends that relation because it drives the dialog forward. IQE, in contrast, treats the user utterance as an unstructured bag of keywords. The unrecognized word "metal" is added to the bag of keywords representing each of the four alternative answers to form four *augmented* queries, and new overlap scores against sentences from SCITEXT are computed. The dialog progresses whenever a new vocabulary word increases the score for the augmented query for the user's chosen answer.

The intuition behind query expansion is that users will explain their answers with salient keywords missing from the original question sentence. The expanded query will overlap with and uprank a support sentence that contains the acquired keywords.

4.2 Performance metrics

Task completion is the proportion of dialogs that end in agreement. Higher task completion indicates that the dialog system is more successful in acquir-

ing enough knowledge by the end of the dialog to change its answer from incorrect to correct.

Dialog length is the number of system and user turns. Shorter dialogs are more efficient.

Acquisition rate is the number of edges in the dKG at the end of each dialog. Acquisition rate measures two contrasting system features:

(1) how much new knowledge is acquired, and

(2) how much *explanatory effort* users expend.

From the perspective of raw knowledge acquisition, higher acquisition rate is better because each dialog adds more edges to the knowledge graph. From the perspective of usability, lower acquisition rate is better provided it doesn't negatively affect dialog success, because it indicates the user is able to successfully correct the system's answer with a fewer number of explanatory relations.

4.3 Results

Our results (Table 1) show both strategies dramatically outperform the baseline and have comparable success rate and dialog length to each other. User-initiative strategies acquire more knowledge per dialog but require more user effort.

	IQE	U.I.	M.I.
Total dialogs	35	27	57
Task completion rate	5.7%	55.6%	49.1%
Mean Dialog Length	14.1	10.6	10.9
Mean acquisition Rate	N/A	13.5	7.4

Table 1: Comparison of knowledge acquisition strategies. Interactive query expansion (IQE)'s poor task completion indicates keywords can't bridge the knowledge gap. Relations are more successful. User-initiative (U.I.) and mixed-initiative (M.I.) strategies have comparable task completion and dialog length, but U.I. extracts twice the relations before getting the correct answer: more knowledge acquired but at the cost of more explanatory effort. User comments indicate M.I. is more satisfying.

We find that the baseline has a very low completion rate of 5%, and longer dialog lengths of 14 turns on average. Interactive query expansion is a poor knowledge acquisition dialog strategy for our task.

In contrast, users were able to successfully correct our system using both strategies about 50% of the time, even though no in-domain ontology guides extractions and no comprehensive dialog model classifies explanations. The average dialog lengths and completion rate for User Initiative (U.I.) and Mixed Initiative (M.I.) strategies was approximately the same, so that choice of strategy had little impact on overall task success. However, strategy has a great effect on acquisition rate. M.I. cuts the knowledge acquisition rate nearly in half when compared to U.I (7.4 novel relations per dialog to 13.5). M.I. learns fewer new relations per dialog with comparable task success, which means each dialog succeeds with much less explanatory effort by the user but also contributes less to the knowledge graph.

User comments indicated that the mixed-initiative strategy was the most enjoyable system to use. We find that open-ended, user-initiative strategies can acquire more helpful relations in a single dialog but guided, mixed-initiative strategies may be more appropriate when usability is taken into account. Because our goal is lifelong interactive knowledge acquisition, the impact of a single dialog on the total knowledge graph is less important than the individual user effort required, and we conclude that the mixed-initiative strategy is preferable.

5 Evaluation of knowledge quality

Experiment 1 evaluated whether users could successfully complete our dialog task. Next, we evaluate whether the total output of our system, all relations acquired during all 431 conducted dialogs, represents useful domain knowledge on this task. We evaluate on questions for which dialogs have been held to investigate whether it's possible to learn any domain knowledge from natural language conversation without a dialog model, irrespective of overfitting. We then use cross-validation to test if knowledge transfers between questions.

As described in section 2, our QA system decomposes each question/answer pair into a true/false statement and chooses as its answer the statement among the four that has the best supporting sentence in a text corpus. Equation (1) scores each question-answer statement by using domain relations to align question concepts to support concepts. The next section describes sources of domain relations.

5.1 Sources of domain knowledge

We compare relations from five sources:

IDENTITY: An edgeless knowledge graph. The only relations are between identical concepts, equivalent to Jaccard overlap of concept keyword roots.

WORDNET: Paraphrase relations from Wordnet. Wordnet (Fellbaum, 1998) is a lexical database of synonyms and hypernyms common in NLP tasks. For example, Snow et al (2006) use Wordnet as training data for ontology induction. To build WORDNET, we draw an edge between every pair of Wordnet concepts (w_s, w_q) for which the Wu-Palmer Similarity (WUP) (Wu and Palmer, 1994) of the first sense in each concept's synset exceeds 0.9, the best-performing WUP threshold we found. Concepts in the Wordnet hierarchy have a higher WUP when they have a closer common ancestor. If a known fact is *Heat energy causes snow to melt*, but a question asks if *ice* melts, then Wordnet should provide the missing knowledge that *ice* acts like *snow*.

PPDB: Paraphrase relations from PPDB (Ganitkevitch et al., 2013) are derived by aligning bilingual parallel texts. PPDB is divided into subsets where the larger subsets have more paraphrases with less precision. We tried all subsets and found the smallest to give the best results, which we report here. The largest performed the worst of all relation sets we tested. We use the lexical paraphrases, which relates unigrams. Concepts are related when at least one concept keyword for each are paraphrases in PPDB. We obtained better performance by stemming PPDB words: for example, if *snows* and *iced* are paraphrases in PPDB then we also considered *snowing* and *icy* to be in PPDB.

KNOWBOT: Each question is answered using relations pooled from all dialogs about all questions. The goal in each dialog is to acquire knowledge helpful to answer the question. If KNOWBOT leads to an increase in QA accuracy over IDENTITY, then we can successfully use open dialog with a human in the loop to learn knowledge that solves a question.

LEAVE-ONE-OUT: Each question is answered only with relations learned during dialogs for every other question. While KNOWBOT uses relations learned from dialogs about the questions on those same questions, LEAVE-ONE-OUT tests whether knowledge generalizes to questions without dialogs. Generalization is only possible when there are at least two questions involving the same concepts. Due to our small number of questions, in the best case we expect only slight improvement.

	%correct
IDENTITY	41%
WORDNET	34%
PPDB	39%
KNOWBOT	**57%**
LEAVE-ONE-OUT	45%

Table 2: QA accuracy on the 107 questions with different sources of domain knowledge. IDENTITY: identity relations only, e.g. "heats" to "heating." WORDNET: Wordnet-derived pseudo-synonyms, e.g. "eagle" to "owl." KNOWBOT: the full, unablated global KG. LEAVE-ONE-OUT: answers each question while ignoring relations acquired during dialogs on that question.

5.2 Results

The results of QA using the different domain knowledge is shown in Table 2. IDENTITY achieves 41% accuracy on this difficult reasoning task, showing that some questions are answerable by searching SCITEXT for supporting sentences with the same concepts as in the question-answer statement. WORDNET works surprisingly poorly. Examination found WORDNET's relations to be of good quality, yet underperform IDENTITY. PPDB performed better but still underperformed IDENTITY. We conclude that general paraphrase bases introduce too much noise to apply directly without manual curation to our science domain, underscoring the need for domain-specific knowledge acquisition.

KNOWBOT achieves accuracy of 57%, a dramatic improvement over both baselines. Importantly, this value does not test *generalization* to unseen questions, since KNOWBOT has held dialogs on these questions. However, it does show that our system can effectively learn about its domain: a poor dialog extraction system will fail to extract any helpful knowledge from users during a training dialog. This is a significant result because it shows that we successfully acquire knowledge to solve many question through conversational interaction without the overhead of a closed dialog model or fixed ontology.

We also tested how well knowledge generalizes with LEAVE-ONE-OUT. Our question set is less suited to evaluate generalization because it covers a wide range of topics with little overlap between

questions. We still found LEAVE-ONE-OUT to be the second-best performer with accuracy of 45%, a 10% relative improvement versus IDENTITY. Redundancy is an effective noise reduction constraint: when LEAVE-ONE-OUT ignores redundancy and includes singleton relations (those originating in a single dialog utterance), its accuracy reduces to 32%.

6 Related work

Knowledge acquisition in dialog has long been a central goal of AI research. Early dialog systems acquired knowledge through ambitious interaction, but were brittle, required hand-defined dialog models and did not scale. Terry Winograd (1972) presented the first dialog system that acquires knowledge about the block world. TEIRESIAS (Davis, 1977) refines inference rules from terse interaction with experts. CONVINCE (Kim and Pearl, 1987) and its prototypes (Leal and Pearl, 1977) learn decision structures through stylized but conversational dialogs. An interactive interface for CYC (Witbrock et al., 2003) learns from experts but don't use natural language. Fernández et al (2011) argue the importance of interactive language learning for conversational agents. Williams et al (2015) combine active learning and dialog to efficiently label training data for dialog act classifiers.

Relatively little work integrates relation extraction and dialog systems. Attribute-value pairs from restaurant reviews can generate system prompts (Reschke et al., 2013), and single-turn exchanges with search engines can populate a knowledge graph (Hakkani-Tur et al., 2014). Dependency relations extracted from individual dialog utterances by a parser also make effective features for dialog act classification (Klüwer et al., 2010).

The work closest to our own, Pappu and Rudnicky (2014a; 2014b), investigates knowledge acquisition strategies for academic events. Their system asks its users open-ended questions in order to elicit information about academic events of interest. They compare strategies by how many new vocabulary words are acquired, so that the best strategy prompts the user to mention the most OOV words. In their most recent work (2014b), they group the acquired researcher names by their interests to form a bipartite graph, and use acquired keywords for query expansion in a simple information retrieval task. Our present contribution builds on this general idea, but we learn an unlimited number of relations and concepts from open dialogs, whereas they learn a small number of relations belonging to a fixed ontology from closed dialogs. We also show the acquired knowledge is objectively useful for QA.

In *closed* dialog systems, the system's dialog model explicitly represents the meaning of every potential user utterance. Any utterance not represented by this comprehensive model is rejected and the user asked to rephrase. Closed dialog systems work well in practice. For example, in the well-studied *slot-filling* or *frame-filling* model, users fill slots to constrain their goal, and an NLU module decomposes user utterances to known actions, slots, and values. A slot-filling system to find flights might map the utterance `U: Show me a flight from Nashville to Seattle on Sunday` to the action *find-flight* and the filled slots *origin = Nashville, destination = Seattle*, and *time = Sunday*. However, for our domain, each distinct question warrants its own actions, slots, and values. Such a complex model would require abundant training data or laboriously handcrafted interpretation rules.

In contrast, an *open* dialog system can usefully interpret, learn from, and respond to user utterances without a comprehensive dialog model. Domain-independent dialog systems with the flexibility to accept novel user utterances are a longstanding goal in dialog research (Polifroni et al., 2003). Recent work to address more open dialog includes bootstrapping a semantic parser from unlabeled dialogs (Artzi and Zettlemoyer, 2011), extracting potential user goals and system responses from backend databases (Hixon and Passonneau, 2013), and inducing slots and slot-fillers from a corpus of human-human dialogs with the use of FrameNet (Chen et al., 2014). These works focus on systems that learn about their domain prior to any human-system dialog. Our system learns about its domain *during* the dialog. While we rely on a limited number of templates to generate system responses, unscripted user utterances can usefully progress the dialog. This allows relation extraction from complex natural language utterances without a closed set of recognized actions and known slot-value decompositions.

7 Discussion and Future Work

KNOWBOT acquires helpful, task-driven relations from conversational dialogs in a difficult QA domain. A dialog is a success when it produces knowledge to solve the question. Extractions increase QA accuracy on questions for which dialogs have been held, indicating that knowledge acquisition dialogs can succeed without a closed dialog model by using task progress and careful pruning to drive natural language understanding. Our method is general enough to scale to any task in which alternative dialog goals can be presented to a user and the system's confidence in each alternative computed from semantic relations between concepts.

Our focus is on facilitated knowledge acquisition rather than question-answering, so we purposefully keep inference simple. The alignment score is a Jaccard overlap modified to use relations, which makes it fast and practical, but results in many ties which we score as incorrect, and also ignores word order. For example, the bag-of-keywords is identical for contradicting answers "changing from liquid to solid" and "changing from solid to liquid." To make this distinction, we could use an alignment score that is sensitive to word order such as an edit distance. We could expand our simple pruning constraints to take more advantage of syntax, for example by using dependency parsers optimized for conversational language (Kong et al., 2014).

The relational model for reasoning is both flexible and powerful (Liu and Singh, 2004). However, in a small number of cases, relations that align known facts with question-answer statements are unlikely to lead to the correct answer. For example, our question set contains a single math problem, *How long does it take for Earth to rotate on its axis seven times? (A) one day (B) one week (C) one month (D) one year*. The multiplication operation necessary to infer the answer from the SCITEXT fact "The Earth rotates, or spins, on its axis once every 24 hours" is not easily represented by our model and requires other techniques (Hosseini et al., 2014).

We observed only slight transfer of knowledge between questions. A larger question set with multiple questions per topic will allow us to better evaluate knowledge transfer. Our long-term goal is learning through any conversational interaction in a completely open domain, but because the fundamental trick that enables model-free NLU is computing progress towards an explicit dialog goal as a function of possible extractions, our current method is limited to tasks with explicit goals.

The simple redundancy filter we use effectively distinguishes salient from noisy relations, but could be improved with a model of relation frequency. We consider all acquired relations equally salient, but future work will examine how to rank relation saliency. We will also examine how dialog features can help distinguish between paraphrase, entailment, and negative relations.

Our open system acquires relations from a wide variety of user explanations without the bottleneck of a hand-built dialog model, but the tradeoff is that we use relatively simple, templated system prompts. However, our collected corpus of real human-system dialogs can be used to improve our system in further iterations. For example, the knowledge graphs we produce are targeted, question-specific semantic networks, which could be used in lieu of FrameNet to induce domain-specific dialog models (Chen et al., 2014). With a dialog model to represent the state space, reinforcement learning could then be employed to optimize our strategies.

While most question-answering systems focus on factoid questions, reasoning tasks such as ours require different techniques. Our method generalizes to other non-factoid QA tasks which could usefully employ relations, such as arithmetic word problems (Hosseini et al., 2014) and biology reading comprehension questions (Berant et al., 2014).

Acknowledgments

This research was conducted at the Allen Institute for Artificial Intelligence. We'd like to thank Luke Zettlemoyer, Mark Yatskar, Rik Koncel-Kedziorski, Eric Gribkoff, Oren Etzioni and the anonymous reviewers for helpful comments, and AI2 interns and colleagues for their support and participation in the user studies. The first author was supported by the National Science Foundation Graduate Research Fellowship Program under Grant Number DGE-1256082. The third author was supported by grants from the Allen Institute for AI (66-9175) and the NSF (IIS-1352249).

References

Yoav Artzi and Luke Zettlemoyer. 2011. Bootstrapping semantic parsers from conversations. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 421–432, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Holger Bast, Debapriyo Majumdar, and Ingmar Weber. 2007. Efficient interactive query expansion with complete search. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 857–860, New York, NY, USA. ACM.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014. Modeling biological processes for reading comprehension. In *Proceedings of EMNLP*.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam Hruschka, and Tom Mitchell. 2010. Toward an architecture for never-ending language learning. In *AAAI Conference on Artificial Intelligence*.

Yun-Nung Chen, William Yang Wang, and Alexander I. Rudnicky. 2014. Leveraging frame semantics and distributional semantics for unsupervised semantic slot induction in spoken dialogue systems. In *2014 IEEE Spoken Language Technology Workshop (SLT 2014)*.

Peter Clark, Niranjan Balasubramanian, Sumithra Bhakthavatsalam, Kevin Humphreys, Jesse Kinkead, Ashish Sabharwal, and Oyvind Tajford. 2014. Automatic construction of inference-supporting knowledge bases. In *4th Workshop on Automated Knowledge Base Construction (AKBC)*.

Randall Davis. 1977. Interactive transfer of expertise: Acquisition of new inference rules. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence. Cambridge, MA, August 1977*, pages 321–328.

Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam Mausam. 2011. Open information extraction: The second generation. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume One*, IJCAI'11, pages 3–10. AAAI Press.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Raquel Fernández, Staffan Larsson, Robin Cooper, Jonathan Ginzburg, and David Schlangen. 2011. Reciprocal learning via dialogue interaction: Challenges and prospects. Proceedings of the IJCAI 2011 Workshop on Agents Learning Interactively from Human Teachers (ALIHT 2011).

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, et al. 2010. Building watson: An overview of the deepqa project. *AI magazine*, 31(3):59–79.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764, Atlanta, Georgia, June. Association for Computational Linguistics.

Dilek Hakkani-Tur, Asli Celikyilmaz, Larry Heck, Gokhan Tur, and Geoff Zweig. 2014. Probabilistic enrichment of knowledge graph entities for relation detection in conversational understanding. In *Proceedings of Interspeech*. ISCA - International Speech Communication Association, September.

Catherine Havasi, Robert Speer, and Jason Alonso. 2007. Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent Advances in Natural Language Processing*, Borovets, Bulgaria, September.

Ben Hixon and Rebecca J. Passonneau. 2013. Open dialogue management for relational databases. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1082–1091, Atlanta, Georgia, June. Association for Computational Linguistics.

Javad Mohammad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. Learning to solve arithmetic word problems with verb categorization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 523–533. Association for Computational Linguistics.

Jin H. Kim and Judea Pearl. 1987. Convince: A conversational inference consolidation engine. *IEEE Trans. Syst. Man Cybern.*, 17(2):120–132, March.

Tina Klüwer, Hans Uszkoreit, and Feiyu Xu. 2010. Using syntactic and semantic based relations for dialogue act recognition. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 570–578, Stroudsburg, PA, USA.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and A. Noah Smith. 2014. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012. Association for Computational Linguistics.

Antonio Leal and Judea Pearl. 1977. An interactive program for conversational elicitation of decision struc-

tures. *IEEE Transactions on Systems, Man, and Cybernetics*, 7(5):368–376.

Hugo Liu and Push Singh. 2004. Commonsense reasoning in and over natural language. In *Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES-2004*. Springer.

Aasish Pappu and Alexander Rudnicky. 2014a. Knowledge acquisition strategies for goal-oriented dialog systems. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 194–198, Philadelphia, PA, U.S.A., June.

Aasish Pappu and Alexander Rudnicky. 2014b. Learning situated knowledge bases through dialog. In *Proceedings of Interspeech*, September.

Joseph Polifroni, Grace Chung, and Stephanie Seneff. 2003. Towards automatic generation of mixed-initiative dialog systems from web content. In *Eurospeech*.

M. F. Porter. 1997. Readings in information retrieval. chapter An Algorithm for Suffix Stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Kevin Reschke, Adam Vogel, and Dan Jurafsky. 2013. Generating recommendation dialogs by extracting information from user reviews. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 499–504, Sofia, Bulgaria, August. Association for Computational Linguistics.

Rion Snow, Daniel Jurafsky, and Andrew Y Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 801–808. Association for Computational Linguistics.

Jason D. Williams, Nobal B. Niraula, Pradeep Dasigi, Aparna Lakshmiratan, Carlos Garcia Jurado Suarez, Mouni Reddy, and Geoff Zweig. 2015. Rapidly scaling dialog systems with interactive learning. In *2015 International Workshop Series on Spoken Dialogue Systems Technology (IWSDS)*, January.

T. Winograd. 1972. *Understanding natural language*. Academic Press.

Michael Witbrock, David Baxter, Jon Curtis, Dave Schneider, Robert Kahlert, Pierluigi Miraglia, Peter Wagner, Kathy Panton, Gavin Matthews, and Amanda Vizedom. 2003. An interactive dialogue system for knowledge acquisition in cyc. In *Proceedings of the IJCAI-2003 Workshop on Mixed-Initiative Intelligent Systems.*, pages 138–145.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics.