Reasoning with Heterogeneous Knowledge for Commonsense Machine Comprehension

Hongyu Lin[1,2] **Le Sun**[1] **Xianpei Han**[1]
[1]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
{hongyu2016,sunle,xianpei}@iscas.ac.cn

Abstract

Reasoning with commonsense knowledge is critical for natural language understanding. Traditional methods for commonsense machine comprehension mostly only focus on one specific kind of knowledge, neglecting the fact that commonsense reasoning requires simultaneously considering different kinds of commonsense knowledge. In this paper, we propose a multi-knowledge reasoning method, which can exploit heterogeneous knowledge for commonsense machine comprehension. Specifically, we first mine different kinds of knowledge (including event narrative knowledge, entity semantic knowledge and sentiment coherent knowledge) and encode them as inference rules with costs. Then we propose a multi-knowledge reasoning model, which selects inference rules for a specific reasoning context using attention mechanism, and reasons by summarizing all valid inference rules. Experiments on RocStories show that our method outperforms traditional models significantly.

1 Introduction

Commonsense knowledge is fundamental in artificial intelligence, and has long been a key component in natural language understanding and human-like reasoning. For example, to understand the relation between sentences "*Mary walked to a restaurant*" and "*She ordered some foods*", we need commonsense knowledge such as "*Mary is a girl*", "*restaurant sells food*", etc. The task of understanding natural language with commonsense knowledge is usually referred as *commonsense machine comprehension*, which has been a hot topic in recent years (Richardson et al., 2013; Weston et al., 2015; Zhang et al., 2016).

Recently, *RocStories* (Mostafazadeh et al., 2016a), a commonsense machine comprehension task, has attached many researchers' attention due to its significant difference from previous machine comprehension tasks. *RocStories* focuses on reasoning with implicit commonsense knowledge, rather than matching with explicit information in given contexts. In this task, a system requires choosing a sentence, namely *hypothesis*, to complete a given commonsense story, called as *premise document*. Table 1 shows two examples. RocStories proposes a challenging benchmark task for evaluating commonsense-based language understanding. As investigated by Mostafazadeh et al.(2016a), this dataset does not have any boundary cases and thus results in 100% human performance.

Commonsense machine comprehension, however, is an natural ability for human but could be very challenging for computers. In general, any world knowledge whatsoever in the reader's mind can affect the choice of an interpretation (Dahlgren et al., 1989). That is, a person can learn any heterogeneous commonsense knowledge and make inference of given information based on all knowledge in his mind. For example, to choose the right hypothesis for the first premise document in Table 1, we needs the event narrative knowledge that "*X does a thorough job*" will lead to "*commends X*", rather than "*fire X*". Besides, people can further confirm their judgement based on the sentimental coherence between "*finish super early*" and "*job well done*". Furthermore, in the second example, even both hypothesises are consistent with the premise document in both event and sentimental facets, we can still infer the right answer easily using the commonsense knowledge that "*puppy*" is a dog, meanwhile "*kitten*" is a cat.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2032–2043
Copenhagen, Denmark, September 7–11, 2017.

Premise Document	Right Hypothesis	Wrong Hypothesis
Ron started his new job as a landscaper today. He loves the outdoors and has always enjoyed working in it. His boss tells him to re-sod the front yard of the mayor's home. Ron is ecstatic, but **does a thorough job** and **finishes super early**.	His boss **commends** him for a **job well done**.	Ron is immediately **fired** for insubordination.
One day, my sister came over to the house to show us her **puppy**. She told us that she had just gotten the puppy across the street. My sons begged me to get them one. I told them that if they would care for it, they could have it.	My son said they would, so we got a **dog**.	We then grabbed a small **kitten**.

Table 1: Examples of RocStories Dataset.

In recent years, many methods have been proposed for commonsense machine comprehension. However, these methods mostly either focus on matching explicit information in given texts (Weston et al., 2014; Wang and Jiang, 2016a,b; Wang et al., 2016b; Zhao et al., 2017), or paid attention to one specific kind of commonsense knowledge, such as event temporal relation (Chambers and Jurafsky, 2008; Modi and Titov, 2014; Pichotta and Mooney, 2016b; Hu et al., 2017) and event causality (Do et al., 2011; Radinsky et al., 2012; Hashimoto et al., 2015; Gui et al., 2016). As discussed above, it is obvious that commonsense machine comprehension problem is far from settled by considering only explicit or a single kind of commonsense knowledge. To achieve human-like comprehension and reasoning, there exist two main challenges:

1) How to mine and represent different kinds of implicit knowledge that commonsense machine comprehension needs. For example, to complete the first example in Table 1, we need a system equipped with the event narrative knowledge that "*commends X*" can be inferred from "*X does a thorough job*", as well as the sentiment coherent knowledge that "*insubordination*" and "*finish super early*" are sentimental incoherent.

2) How to reason with various kinds of commonsense knowledge. As shown above, knowledge that reasoning process needs varies for different contexts. For human-like commonsense machine comprehension, a system should take various kinds of knowledge into consideration, decide what knowledge will be utilized in a specific reasoning contexts, and make the final decision by taking all utilized knowledge into consideration.

To address the above problems, this paper proposes a new commonsense reasoning approach, which can mine and exploit heterogeneous knowledge for commonsense machine comprehension. Specifically, we first mine different kinds of knowledge from raw text and relevant knowledge base, including event narrative knowledge, entity semantic knowledge and sentiment coherent knowledge. These heterogeneous knowledge are encoded into a uniform representation – inference rules between elements under different kinds of relations, with an inference cost for each rule. Then we design a rule selection model using attention mechanism, modeling which inference rules will be applied in a specific reasoning context. Finally, we propose a multi-knowledge reasoning model, which measures the reasoning distance from a premise document to a hypothesis as the expected cost sum of all inference rules applied in the reasoning process.

By modeling and exploiting heterogeneous knowledge during commonsense reasoning, our method can achieve more accurate and more robust performance than traditional methods. Furthermore, our method is a general framework, which can be extended to incorporate new knowledge easily. Experiments show that our method achieves a 13.7% accuracy improvement on the standard RocStories dataset, a significant improvement over previous work.

2 Commonsense Knowledge Acquisition for Machine Comprehension

As described above, various knowledge can be exploited for machine comprehension. In this section, we describe how to mine different knowledge from different sources. Specifically, we mine three types of commonly used commonsense knowledge, including: 1)*Event narrative knowledge*, which captures temporal and causal relations between events; 2)*Entity semantic knowledge*, which captures semantic relations between entities; 3)*Sentiment coherent knowledge*, which captures sentimental coherence between elements.

In this paper, we represent commonsense knowledge as a set of inference rules given in the form of $X \xrightarrow{f} Y : s$, which means that element Y can be inferred from element X under relation f, with an inference cost s. An element can stand

	Antecedent	Consequent	Relation	Cost
①	Mary	she	coreference	0.0
②	restaurant	order	narrative	0.1
③	restaurant	food	associative	0.1
④	restaurant	food	narrative	0.3
⑤	Mary	order	narrative	0.5
⑥	walk	sleep	narrative	0.8
⑦	walk	food	narrative	0.9

Table 2: Examples of Inference Rules.

for either event, entity or sentiment, and this paper represents elements using lemmatized nouns, verbs and adjectives. The lexical element representation can also be easily extended to structural representation, like the one in (Chambers and Jurafsky, 2008), if needed. However, in auxiliary experiments we found that using structural elements results in severe sparseness and noises which in turn will hurt the reasoning performance. Therefore, we think an individual work is needed to solve it. Table 2 demonstrates several examples of inference rules. In following, we describe how to mine different types of inference rules.

2.1 Mining Event Narrative Knowledge

Event narrative knowledge captures structured temporal and casual knowledge about stereotypical event sequences, which is fundamental for commonsense machine comprehension. For example, we can infer "X ordered some foods" from "X walked to a restaurant" using event narrative knowledge. Previous work (Chambers and Jurafsky, 2008; Rudinger et al., 2015) proves that event narrative knowledge can be mined from raw texts unsupervisedly. So we propose two models to encode this knowledge using inference rules.

The first one is based on ordered PMI, which is also proposed by Rudinger et al. (2015). Given two element e_1 and e_2, this model calculates the cost of inference rule $e_1 \xrightarrow{narrative} e_2$ as:

$$cost(e_1 \rightarrow e_2) = -log\frac{C(e_1, e_2)}{C(e_1, *), C(*, e_2)} \quad (1)$$

Here $C(e_1, e_2)$ is the order sensitive count that element e_1 occurs before element e_2 in different sentences of the same document.

The second model is a variant of the skip-gram model (Mikolov et al., 2013). The goal of this model is to find element representations which can accurately predict relevant elements in sentences afterwards. Formally, given n asymmetric pairs of elements $(e_1^1, e_2^1), (e_1^2, e_2^2), ..., (e_1^n, e_2^n)$ identified from training data, the objective of our model is to maximize the average log probability $\frac{1}{n}\sum_{i=1}^{n} logP(e_2^i|e_1^i)$. And the probability $P(e_2|e_1)$ is defined using the softmax function:

$$P(e_2|e_1) \propto exp({\mathbf{v}'_{e_2}}^T \mathbf{v}_{e_1}) \quad (2)$$

where $\mathbf{v}_e$ and $\mathbf{v}'_e$ are "antecedent" and "consequent" vector representation of element e, respectively. We use the negative inner product $-{\mathbf{v}'_{e_2}}^T \mathbf{v}_{e_1}$ as the cost of inference rule $e_1 \xrightarrow{skip-gram} e_2$.

2.2 Mining Entity Semantic Knowledge

Entities, often serving as event participants or environment variables, are important components of commonsense stories. Intuitively, an entity in hypothesis is reasonable if we can identify semantic relations between it and some parts of premise document. For example, if a premise document contains "Starbucks", then "coffeehouse" and "latte" will be reasonable entities in hypothesis since "Starbucks" is a possible coreference of "coffeehouse" and it is semantically related to "latte".

Specifically, we identify mainly two kinds of semantic relations between entities for commonsense machine comprehension:

1) *Coreference relation,* which indicates that two elements refer to the same entity in environment. In stories, besides to pronouns, an entity is often referred using its hypernyms, e.g, the second example in Table 1 uses "dog" to refer to "puppy". Motivated by this observation, we mine coreference knowledge between elements using Wordnet (Kilgarriff and Fellbaum, 2000): $X \xrightarrow{coref} Y$ is an inference rule with cost 0 if X and Y are lemmas in the same Wordnet synset, or with hyponymy relation in Wordnet. Otherwise, the cost of inference rules between this element-pair under this relation will be 1.

2) *Associative relation,* which captures the semantic relatedness between two entities, i.e., "starbucks" → "latte", "restaurant" → "food", etc. This paper mines associative relations between entities from Wikipedia[1], using the method proposed by Milne and Witten(2008). Specifically, given two entities e_1 and e_2, we compute the semantic distance $dist(e_1, e_2)$ between them as:

$$dist(e_1, e_2) = \frac{log(max(|E_1|, |E_2|) - log(|E_1 \bigcap E_2|))}{log(|W|) - log(min(|E_1|, |E_2|))} \quad (3)$$

where E_1 and E_2 are the sets of all entities that link to these two entities in Wikipedia respectively,

[1]https://www.wikipedia.org/

and W is the entire Wikipedia. We set the cost of inference rule $e_1 \xrightarrow{associative} e_2$ as $dist(e_1, e_2)$.

2.3 Mining Sentiment Coherent Knowledge

Sentiment is one of the central and pervasive aspects of human experience (Ortony et al., 1990). It plays an important role in commonsense stories, i.e., a reasonable hypothesis should be sentimental coherent with its premise document. In this paper, we mine sentiment coherence rules using SentiWordnet (Baccianella et al., 2010), in which each synset of Wordnet is assigned with three sentiment scores: positivity, negativity and objectivity.

Concretely, to identify sentimental coherence rule between two element e_1 and e_2, we first compute the positivity, negativity and objectivity scores of every element by averaging the scores of all synsets it's in, then we identify an element to be subjective if its objectivity score is smaller than a threshold, and the distance between its positivity and negativity score is greater than a threshold. Finally, for an inference rule $e_1 \xrightarrow{senti} e_2$, we set its cost to 1 if e_1 and e_2 are both subjective and have opposite sentimental polarity, to -1 if they are both subjective and their sentimental polarity are the same, and to 0 for other cases. For example, we will mine inference rules "good $\xrightarrow{senti}$ happy : -1", "perfect $\xrightarrow{senti}$ sad : 1" and "young $\xrightarrow{senti}$ happy : 0".

2.4 Metric Learning to Calibrate Cost Measurement

So far, we have extracted many inference rules under different relations. However, because we extract them from different sources and estimate their costs using different measurements, the cost metrics of these rules may not be consistent with each other. To exploit different types of inference rules in a unified framework, we here propose a metric learning based method to calibrate their costs.

Given an input distance function, a metric learning method constructs a new distance function which is "better" than the original one with supervision regarding an ideal distance (Kulis, 2012). To calibrate inference rule cost, we add a nonlinear layer to the original cost s_r of inference rule r under relation f:

$$c_r = sigmoid(w_f s_r + b_f) \quad (4)$$

Here c_r is the metric-unified inference cost of inference rule r, w_f and b_f are calibration parameters for inference rules of relation f. We use sigmoid function in order to normalize costs into 0 to 1. Calibration parameters will be trained along with other parameters in our model. See Section 3.4 for detail.

2.5 Dealing with Negation

One important linguistic phenomenon needs to specifically consider is negation. Here we discuss how to solve negation in our model.

We use $\neg X$ to represent an element X modified by a negation word (the existence of negation is detected using dependency relations). Under event narrative relation and sentiment coherent relation, the existence of negation will reverse the conclusion. So we add three additional negation related inference rules for rule $X \xrightarrow{f} Y : s$ under these relations, including $\neg X \xrightarrow{f} Y : 1 - s$, $X \xrightarrow{f} \neg Y : 1 - s$ and $\neg X \xrightarrow{f} \neg Y : s$. Here s is the calibrated cost of the original inference rule. For entity semantic relations, we just ignore the negation since it will not affect the inference under these relations.

3 Machine Comprehension via Commonsense Reasoning

This section describes how to leverage acquired knowledge for commonsense machine comprehension. We first define how to infer from a premise document to a hypothesis using inference rules. Then we model how to choose inference rules for a specific reasoning context. Finally, we describe how to measure the reasoning distance from a premise document to a hypothesis by summarizing the costs of all possible inferences.

3.1 Inference from Premise Document to Hypothesis

Given a premise document $D = \{d_1, d_2, ..., d_m\}$ containing m elements, a hypothesis $H = \{h_1, h_2, ..., h_n\}$ containing n elements, a valid inference R from D to H is a set of inference rules that all elements in H can be inferred from one element in D using one and only one rule in R. This definition means that all elements in H should be covered by consequents of inference rules in R, as well as all antecedents of inference rules in R should come from D. Figure 1 shows some inference examples, where (a), (b) and (d) are valid inferences, but (c) is not a valid inference because its rules can not cover all elements in hypothesis.

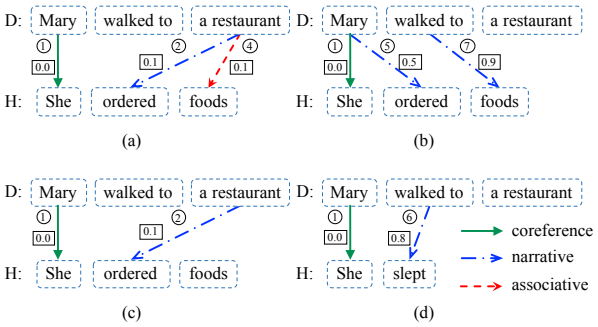

Figure 1: Examples of inferences. Numbers in circle indicates the proposed inference rules in Table 2, and values in rectangle are their costs.

By the definition, the size of R and the size of H are equal. So we use r_i to denote the inference rule in R that applied to derive element h_i in H, i.e., $R = \{r_1, r_2, ..., r_n\}$.

Based on the above definition, we can naturally define the cost of an inference R as the cost sum of all inference rules in R. In Figure 1, the cost for inference (a) is $0.0 + 0.1 + 0.1 = 0.2$, and for inference (d) is $0.0 + 0.8 = 0.8$.

3.2 Modeling Inference Probability using Attention Mechanism

Obviously, there exist multiple valid inferences for a premise document and a hypothesis. For example, in Figure 1, both (a) and (b) are valid inferences for the same premise document and hypothesis. To identify whether a hypothesis is reasonable, we need to consider all possible inferences. However, in human reasoning process, not all inference rules have the same possibility to be applied, because the more reasonable inference will be proposed more likely. In Figure 1, inference (a) should have a higher probability than inference (b) because it is more reasonable to infer "*foods*" from "*a restaurant*" with *associative relation*, rather than from "*walked to*" with *narrative relation*. Besides, the possibility of proposing an inference should not depend on its cost, e.g., inference (d) should have high possibility to be proposed despite its high cost, because we often infer event "*sleep*" from another event using inference rules under *narrative relation*. As examples mentioned above, the "cost" measures the "correctness" of an inference rule. A rule with low cost is more likely to be "reasonable", and a rule with high cost is more likely to be a contradiction with commonsense. On the other hand, the "possibility" should measure how likely a rule will be applied in a given context, which does not depend on the "cost" but on the nature of the rule and the given context. Motivated by above observations, we endow each inference a probability $P(R|D, H)$, indicating the possibility that R is chosen to infer hypothesis H from premise document D. For simplicity, we assume that each element in hypothesis is independently inferred using individual inference rule, then $P(R|D, H)$ can be written as:

$$P(R|D,H) = \prod_{i=1}^{n} P(r_i|D,H) \quad (5)$$

$$= \prod_{i=1}^{n} P(r_i|D,h_i) \quad (6)$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{m} P(r_i, d_j|D,h_i) \quad (7)$$

Equation (7) clearly shows how an inference rule is selected given the premise document D and the element h_i in hypothesis. It depends on which element d_j in D will be selected and which relation f will be used to infer h_i from d_j. We then refactor the probability $P(r_i, d_j|D, h_i)$ to be:

$$P(r_i, d_j|D, h_i) = \begin{cases} 0 & , antecedent(r_i) \neq d_j \\ g(h_i, d_j, f(r_i); D) & , otherwise \end{cases} \quad (8)$$

Here $f(r)$ is the relation type of inference rule r, and $g(h, d, f; D)$ is defined as:

$$g(h,d,f;D) = \frac{s(h,d)a(h,f)a(d,f)}{\sum_{f\in\mathcal{F}}\sum_{d\in D} s(h,d)a(h,f)a(d,f)} \quad (9)$$

Here $\mathcal{F}$ denotes all relation types of inference rules, $s(e_1, e_2)$ is a matching function between two elements e_1 and e_2, measuring by cosine similarity based on GoogleNews word2vec (Mikolov et al., 2013). And $a(e, f)$ is an attention function measuring how likely an element e will be involved with rules under relation f:

$$a(e,f) = \mathbf{v_f}^T tanh(\mathbf{W_f e} + \mathbf{b_f}) \quad (10)$$

where $\mathbf{v_f} \in \mathbb{R}^K$, $\mathbf{W_f} \in \mathbb{R}^{K\times F}$ and $\mathbf{b_f} \in \mathbb{R}^K$ are attention parameters of relation f, and $\mathbf{e} \in \mathbb{R}^F$ is the feature vector of element e. Here K is the size of attention hidden layer and F is the dimension of feature vector. We consider three types of features, as shown in Table 3. Using attention mechanism, our method models the possibility that an inference rule is applied during the inference from a premise document to a hypothesis by considering the relatedness between elements and knowledge category, as well as the relatedness between two elements, which make it able to select the most reasonable inference rules to derive each part of the hypothesis.

Feature	Description
Syntax Features	
is_verb	whether this element is a verb
is_noun	whether this element is a noun
is_adj	whether this element is an adjective
Lexical Features	
is_event	whether this element belongs to hyponymy of *event sysnset* in Wordnet
is_entity	whether this element belongs to hyponymy of *physical_entity sysnset*
is_attr	whether this element belongs to hyponymy of *attribute sysnset*
named_entity	the named entity type of this element
Semantic Features	
word embeddings	300 dimension embeddings of GoogleNews word2vec model

Table 3: Features for element-relation attention.

3.3 Reasoning Distance Between Premise Document and Hypothesis

Given a premise document, this section shows how to measure whether a hypothesis is coherent using above inference model. Given all valid inferences from D to H and the probability $P(R|D,H)$ of selecting inference R to infer H from D, we measure the reasoning distance $L(D \rightarrow H)$ as the expected cost sum of all valid inferences:

$$L(D \rightarrow H) = E_{P(R|D,H)}[cost(R)] \quad (11)$$

$$= E_{P(R|D,H)}[\sum_{i=1}^{n} cost(r_i)] \quad (12)$$

Then using Equation (6) and Equation (7), we can further rewrite the equation into:

$$L(D \rightarrow H) = \sum_{R}[\prod_{i=1}^{n} P(r_i|D,h_i)] \cdot [\sum_{i=1}^{n} cost(r_i)] \quad (13)$$

$$= \sum_{i=1}^{n} P(r_i|D,h_i) \cdot cost(r_i) \quad (14)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} P(r_i,d_j|D,h_i) \cdot cost(r_i) \quad (15)$$

Equation (15) shows that in our framework, the final cost of inferring the element h_i in the hypothesis is the expected cost of all valid inference rules which can derive h_i from one element in the premise document.

3.4 Model Learning

Following Huang et al. (2013), our model measures the posterior probability of choosing hypothesis H as the answer of premise document D through a softmax function:

$$P(H|D) = \frac{exp(-\gamma L(D \rightarrow H))}{\sum_{H' \in \mathcal{H}_D} exp(-\gamma L(D \rightarrow H'))} \quad (16)$$

Here $\mathcal{H}_D$ is all candidate hypothesises for D, and γ is a positive smoothing factor. We train our model by maximizing the likelihood of choosing right hypothesis H^+ for D:

$$\mathcal{L}(\theta) = -log \prod_{(D,H^+)} P(H^+|D) \quad (17)$$

where θ is the parameter set of our model, including calibration parameters in Section 2.4 and attention parameters in Section 3.2. $\mathcal{L}(\theta)$ is differentiable so we can estimate θ using any gradient-based optimization algorithm.

4 Experiments

4.1 Experimental Settings

Data Preparation. We evaluated our approach on the *Test Set Spring 2016* of RocStories, which consists of 1871 commonsense stories, with each story has two candidate story endings. Because stories in the training set of RocStories do not contain wrong hypothesis, and our model has a compact size of parameters, we estimated the parameters of our model using the *Validation Set Spring 2016* of RocStories with 1871 commonsense stories.

We mined event narrative knowledge from the *Training Set Spring 2016* of RocStories, which consists of 45502 commonsense stories. We performed lemmatisation, part of speech annotation, named entity tagging, and dependency parsing using Stanford CoreNLP toolkits (Manning et al., 2014). We used the Jan. 30, 2010 English version of Wikipedia and processed it according to the method described by Hu et al. (2008).

Model Training. We used normalized initialization (Glorot and Bengio, 2010) to initialize attention parameters in our model. For calibration parameters, we initialized all $\mathbf{w_f}$ to 1 and $\mathbf{b_f}$ to 0. The model parameters were trained using mini-batch stochastic gradient descent algorithm. As for hyper-parameters, we set the batch size as 32, the learning rate as 1, the dimension of attention hidden layer K as 32, and the smoothing factor γ as 0.5.

Baselines. We compared our approach with following three baselines:

1) **Narrative Event Chain** (Chambers and Jurafsky, 2008), which scores hypothesis using PMI scores between events. We used a simplified version of the original model by using only verbs as event, ignoring the dependency relation between verbs and their participants. We found such a simplified version achieved better performance than its original one whose performance was reported in (Mostafazadeh et al., 2016a).

2) **Deep Structured Semantic Model (DSS-**

M) (Huang et al., 2013), which achieved the best performance on RocStories as reported by Mostafazadeh et al.(2016a). This model measures the reasoning score between a premise document D and a hypothesis H by calculating the cosine similarity between the overall vector representations of D and H, and do not consider any other task-relevant knowledge.

3) **Recurrent Neural Network(RNN) Model** proposed by Pichotta and Mooney(2015), which transforms all events and their arguments into a sequence and predict next events and arguments using a Long Short-Term Memory network. We used the average generating probability of all elements in H as the reasoning score, and choose the hypothesis with largest reasoning score as the system answer.

4.2 Overall Performance

System	Accuracy
Narrative Event Chain	57.62%
DSSM	58.52%
RNN Model	58.93%
Our Model	**67.02%**

Table 4: Comparison of accuracy for our model and three baselines on RocStories Spring 2016 Test Set. The result of DSSM is adapted from (Mostafazadeh et al., 2016a).

Table 4 shows the results. From this table, we can see that:

1) Our model outperforms all baselines significantly. Compared with baselines, the accuracy improvement on test set is at least 13.7%. This demonstrates the effectiveness of our model by mining and exploiting heteregenous knowledge.

2) The event narrative knowledge only is insufficient for commonsense machine comprehension. Compared with Narrative Event Chain Model, our model achieves a 16.3% accuracy improvement by considering richer commonsense knowledge, rather than only narrative event knowledge.

3) It is necessary to distinguish different kinds of commonsense relations for machine comprehension and commonsense reasoning. Compared with DSSM and RNN, which model all relations between two elements using a single semantic similarity score, our model achieves significant accuracy improvements by modeling, distinguishing and selecting different types of commonsense relations between different kinds of elements.

4.3 Effects of Different Knowledge

To investigate the effect of different kinds of knowledge in our model, we conducted two groups of experiments.

The first group of experiments was conducted using only one kind of knowledge at a time in our model. Table 5 shows the results. We can see that using a single kind of knowledge is insufficient for commonsense machine comprehension: all single-knowledge settings cannot achieve competitive performance to the all-knowledge setting.

System	Accuray
Event Narrative Knowledge	60.98%
Entity Semantics Knowledge	57.14%
Sentiment Coherent Knowledge	61.30%
Our Model(All Knowledge)	**67.02%**

Table 5: Comparison of the performance using single type of knowledge.

The second group of experiments was conducted to investigate whether different knowledge can complement each other. We conducted experiments by removing one kind of knowledge from our final model at a time, and investigate the change of accuracy.

System	Accuracy
Our Model(All Knowledge)	**67.02%**
-w/o Event Narrative Knowledge	63.65%
-w/o Entity Semantic Knowledge	64.89%
-w/o Sentiment Coherent Knowledge	62.85%

Table 6: Comparison of the performance by removing one single type of knowledge.

Table 6 shows the results. We can find that removing any kind of knowledge will reduce the accuracy. This verified that all kinds of knowledge containing unique complementary information, which cannot be covered by other types of knowledge.

4.4 Effect of Inference Probability

This section investigates the effect of inference rule selection probability, and whether our attention mechanism can effectively model the possibility of inference rule selection. We compared our method with following two heuristic settings:

1) **Minimum Cost Mechanism**, which measures the reasoning distance by only selecting the inference rule with minimum cost for each hypothesis element.

2) **Average Cost Mechanism**, which measures the reasoning distance by setting equal probabilities to all inference rules that can infer a hypothesis element from a premise document element.

System	Accuracy
Minimum Cost Mechanism	54.84%
Average Cost Mechanism	63.01%
Our Model(Attention Mechanism)	**67.02%**

Table 7: Comparison of the performance using different inference rule selection mechanism.

Table 7 show the results. We can see that: 1) the minimum cost mechanism cannot achieve competitive performance, we believe this is because the selection of rules should not depend on the cost of them, and considering all valid inferences is critical for reasoning; 2) our attention mechanism can effectively model the inference rule selection possibility. Compared with the average cost mechanism, our method achieved a 6.36% accuracy improvement. This also verified the necessity of an effective inference rule probability model.

4.5 Effect of Negation Rules

This section investigates the effect of special handling of negation mentioned in Section 2.5. To investigate the necessity of negation rules proposed in our model, we conducted experiments by removing all negation rules from original system, and investigate the change of accuracy.

System	Accuracy
Our Model	**67.02%**
-w/o Negation Rules	63.12%

Table 8: Comparison of the performance by removing negation rules.

Table 8 show the results. We can see that removing negation rules will significantly drop the system performance, which confirm the effectiveness of our proposed negation rules.

5 Related Work

Endowing computers with the ability of understanding commonsense story has long a goal of natural language processing. There exist two big challenges: 1)Matching explicit information in the given context; 2)Incorporating implicit commonsense knowledge into human-like reasoning process. Previous machine comprehension tasks (Richardson et al., 2013; Weston et al., 2015; Hermann et al., 2015; Rajpurkar et al., 2016) mainly focus on the first challenge, leading their solutions focusing on semantic matching between texts (Weston et al., 2014; Kumar et al., 2015; Narasimhan and Barzilay, 2015; Smith et al., 2015; Sukhbaatar et al., 2015; Hill et al., 2015; Wang et al., 2015, 2016a; Cui et al., 2016; Trischler et al., 2016a,b; Kadlec et al., 2016; Kobayashi et al., 2016; Wang and Jiang, 2016b), but ignore the second issues. One notable task is SNLI (Bowman et al., 2015), which considers entailment between two sentences. This task, however, only provides shallow context and thus needs a few kinds of implicit knowledge (Rocktäschel et al., 2015; Wang and Jiang, 2016a; Angeli et al., 2016; Wang et al., 2016b; Parikh et al., 2016; Henderson and Popa, 2016; Zhao et al., 2017).

Realizing that story understanding needs commonsense knowledge, many researches have been proposed to learn structural event knowledge. Chambers and Jurafsky (2008) first proposed an unsupervised approach to learn partially ordered sets of events from raw text. Many expansions have been introduced later, including unsupervisedly learning narrative schemas and scripts (Chambers and Jurafsky, 2009; Regneri et al., 2011), event schemas and frames (Chambers and Jurafsky, 2011; Balasubramanian et al., 2013; Sha et al., 2016; Huang et al., 2016; Mostafazadeh et al., 2016b), and some generative models to learn latent structures of event knowledge (Cheung et al., 2013; Chambers, 2013; Bamman et al., 2014; Nguyen et al., 2015). Another direction for learning event-centred knowledge is causality identification (Do et al., 2011; Radinsky et al., 2012; Berant et al., 2014; Hashimoto et al., 2015; Gui et al., 2016), which tried to identify the causality relation in text.

For reasoning over these knowledge, Jans et al. (2012) extend introduced skip-grams for collecting statistics. Further improvements include incorporating more information and more complicated models (Radinsky and Horvitz, 2013; Modi and Titov, 2014; Ahrendt and Demberg, 2016). Recent researches tried to solve event prediction problem by transforming it into an language modeling paradigm (Pichotta and Mooney, 2014, 2015, 2016a,b; Rudinger et al., 2015; Hu et al., 2017).

The principal difference between previous work and our method is that we not only take various kinds of implicit commonsense knowledge into consideration, but also provide a highly

extensible framework to exploit these kinds of knowledge for commonsense machine comprehension. We also notice the recent progress in RocStories (Mostafazadeh et al., 2017). Rather than inferring a possible ending generated from document, recent systems solve this task by discriminatively comparing two candidates. This enables very strong stylistic features being added explicitly (Schwartz et al., 2017; Bugert et al., 2017) or implicitly (Schenk and Chiarcos, 2017), which can select hypothesis without any consideration of given document. Also, some augmentation strategies are introduced to produce more training data (Roemmele and Gordon, 2017; Mihaylov and Frank, 2017; Bugert et al., 2017). These methods are dataset-sensitive and are not the main concentration of our paper.

6 Conclusions and Future Work

This paper proposes a commonsense machine comprehension method, which performs effective commonsense reasoning by taking heterogeneous knowledge into consideration. Specifically, we mine commonsense knowledge from heterogeneous knowledge sources and simultaneously exploit them by proposing a highly extensible multi-knowledge reasoning framework. Experiment results shown that our method surpasses baselines by a large margin.

Currently, there are little labeled training instances for commonsense machine comprehension, for future work we want to address this issue by developing semi-supervised or unsupervised approaches.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grants no. 61433015 and 61572477, the National High Technology Development 863 Program of China under Grants no. 2015AA015405, and the Young Elite Scientists Sponsorship Program no. YESS20160177. Moreover, we sincerely thank the reviewers for their valuable comments.

References

Simon Ahrendt and Vera Demberg. 2016. Improving event prediction by representing script participants. In *Proceedings of NAACL-HLT*, pages 546–551.

Gabor Angeli, Neha Nayak, and Christopher D Manning. 2016. Combining natural logic and shallow reasoning for question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 442–452.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204.

Niranjan Balasubramanian, Stephen Soderland, Oren Etzioni Mausam, and Oren Etzioni. 2013. Generating coherent event schemas at scale. In *EMNLP*, pages 1721–1731.

David Bamman, Brendan O'Connor, and Noah A Smith. 2014. Learning latent personas of film characters. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, page 352.

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 *EMNLP*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Michael Bugert, Yevgeniy Puzikov, R Andreas, Judith Eckle-kohler, Teresa Martin, and Eugenio Mart. 2017. LSDSem 2017 : Exploring Data Generation Methods for the Story Cloze Test. In *The 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (ISDSEM 2017)*, 2016, pages 56–61, Valencia, Spain.

Nathanael Chambers. 2013. Event schema induction with a probabilistic entity-driven model. In *EMNLP*, volume 13, pages 1797–1807.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 602–610. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2011. Template-based information extraction without the templates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 976–986. Association for Computational Linguistics.

Nathanael Chambers and Daniel Jurafsky. 2008. Unsupervised learning of narrative event chains. In *ACL*, volume 94305, pages 789–797. Citeseer.

Jackie Chi Kit Cheung, Hoifung Poon, and Lucy Vanderwende. 2013. Probabilistic frame induction. *arXiv preprint arXiv:1302.4813*.

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2016. Attention-over-attention neural networks for reading comprehension. *arXiv preprint arXiv:1607.04423*.

Kathleen Dahlgren, Joyce McDowell, and Edward P Stabler. 1989. Knowledge representation for commonsense reasoning with text. *Computational Linguistics*, 15(3):149–170.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 294–303. Association for Computational Linguistics.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, pages 249–256.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. 2016. Event-driven emotion cause extraction with corpus construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1639–1649.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, and Jong-Hoon Oh. 2015. Generating event causality hypotheses through semantic relations. In *AAAI*, pages 2396–2403.

James Henderson and Diana Nicoleta Popa. 2016. A vector space for distributional semantics for entailment. *arXiv preprint arXiv:1607.03780*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Jian Hu, Lujun Fang, Yang Cao, Hua-Jun Zeng, Hua Li, Qiang Yang, and Zheng Chen. 2008. Enhancing text clustering by leveraging wikipedia semantics. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 179–186. ACM.

Linmei Hu, Juanzi Li, Liqiang Nie, Xiao-Li Li, and Chao Shao. 2017. What happens next? future subevent prediction using contextual hierarchical lstm. In *Proceedings of the 31th AAAI Conference on Artificial Intelligence*.

Lifu Huang, T Cassidy, X Feng, H Ji, CR Voss, J Han, and A Sil. 2016. Liberal event extraction and event schema induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-16)*.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 2333–2338. ACM.

Bram Jans, Steven Bethard, Ivan Vulić, and Marie Francine Moens. 2012. Skip n-grams and ranking functions for predicting script events. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 336–344. Association for Computational Linguistics.

Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. *arXiv preprint arXiv:1603.01547*.

Adam Kilgarriff and Christiane Fellbaum. 2000. Wordnet: An electronic lexical database.

Sosuke Kobayashi, Ran Tian, Naoaki Okazaki, and Kentaro Inui. 2016. Dynamic entity representation with max-pooling improves machine reading. In *Proceedings of NAACL-HLT*, pages 850–855.

Brian Kulis. 2012. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364.

Ankit Kumar, Ozan Irsoy, Jonathan Su, James Bradbury, Robert English, Brian Pierce, Peter Ondruska, Ishaan Gulrajani, and Richard Socher. 2015. Ask me anything: Dynamic memory networks for natural language processing. *CoRR, abs/1506.07285*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Todor Mihaylov and Anette Frank. 2017. Story Cloze Ending Selection Baselines and Data Examination. In *The 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (ISDSEM 2017)*, pages 2–7, Valencia, Spain.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

David Milne and Ian H Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM.

Ashutosh Modi and Ivan Titov. 2014. Inducing neural models of script knowledge. In *CoNLL*, volume 14, pages 49–57.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016a. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of NAACL-HLT*, pages 839–849.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016b. Caters: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the The 4th Workshop on EVENTS: Definition, Detection, Coreference, and Representation, San Diego, California, June. Association for Computational Linguistics*.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael William Chambers, and James F. Allen. 2017. LSDSem 2017 Shared Task : The Story Cloze Test. In *The 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (ISDSEM 2017)*, 2016, pages 1–5, Valencia, Spain.

Karthik Narasimhan and Regina Barzilay. 2015. Machine comprehension with discourse relations. In *ACL (1)*, pages 1253–1262.

Kiem-Hieu Nguyen, Xavier Tannier, Olivier Ferret, and Romaric Besançon. 2015. Generative event schema induction with entity disambiguation. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics (ACL-15)*.

Andrew Ortony, Gerald L Clore, and Allan Collins. 1990. *The cognitive structure of emotions*. Cambridge university press.

Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*.

Karl Pichotta and Raymond J Mooney. 2014. Statistical script learning with multi-argument events. In *EACL*, volume 14, pages 220–229.

Karl Pichotta and Raymond J Mooney. 2015. Learning statistical scripts with lstm recurrent neural networks. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*.

Karl Pichotta and Raymond J Mooney. 2016a. Statistical script learning with recurrent neural networks. *EMNLP 2016*, page 11.

Karl Pichotta and Raymond J Mooney. 2016b. Using sentence-level lstm language models for script inference. *arXiv preprint arXiv:1604.02993*.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *Proceedings of the 21st international conference on World Wide Web*, pages 909–918. ACM.

Kira Radinsky and Eric Horvitz. 2013. Mining the web to predict future events. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 255–264. ACM.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Michaela Regneri, Alexander Koller, Josef Ruppenhofer, and Manfred Pinkal. 2011. Learning script participants from unlabeled data. In *RANLP*, pages 463–470.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*, volume 3, page 4.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

Melissa Roemmele and Andrew M Gordon. 2017. An RNN-based Binary Classifier for the Story Cloze Test. In *The 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (ISDSEM 2017)*, pages 74–80, Valencia, Spain.

Rachel Rudinger, Pushpendre Rastogi, Francis Ferraro, and Benjamin Van Durme. 2015. Script induction as language modeling. In *EMNLP*, pages 1681–1686.

Niko Schenk and Christian Chiarcos. 2017. Resource-Lean Modeling of Coherence in Commonsense Stories. In *The 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (ISDSEM 2017)*, pages 68–73, Valencia, Spain.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, Noah A Smith, and Computer Science. 2017. Story Cloze Task : UW NLP System. In *The 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (ISDSEM 2017)*, pages 52–55, Valencia, Spain.

Lei Sha, Sujian Li, Baobao Chang, and Zhifang Sui. 2016. Joint learning templates and slots for event schema induction. In *Proceedings of NAACL-HLT*, pages 428–434.

Ellery Smith, Nicola Greco, Matko Bosnjak, and Andreas Vlachos. 2015. A strong lexical matching method for the machine comprehension test. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1693–1698. Association for Computational Linguistics.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.

Adam Trischler, Zheng Ye, Xingdi Yuan, Jing He, Phillip Bachman, and Kaheer Suleman. 2016a. A parallel-hierarchical model for machine comprehension on sparse data. *arXiv preprint arXiv:1603.08884*.

Adam Trischler, Zheng Ye, Xingdi Yuan, and Kaheer Suleman. 2016b. Natural language comprehension with the epireader. *arXiv preprint arXiv:1606.02270*.

Bingning Wang, Shangmin Guo, Kang Liu, Shizhu He, and Jun Zhao. 2016a. Employing external rich knowledge for machine comprehension. In *Proceedings of IJCAI*.

Hai Wang, Mohit Bansal, Kevin Gimpel, and David A McAllester. 2015. Machine comprehension with syntax, frames, and semantics. In *ACL (2)*, pages 700–706.

Shuohang Wang and Jing Jiang. 2016a. Learning natural language inference with lstm. In *Proceedings of NAACL-HLT*, pages 1442–1451.

Shuohang Wang and Jing Jiang. 2016b. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905*.

Zhiguo Wang, Haitao Mi, and Abraham Ittycheriah. 2016b. Sentence similarity learning by lexical decomposition and composition. *arXiv preprint arXiv:1602.07019*.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2016. Ordinal common-sense inference. *arXiv preprint arXiv:1611.00601*.

Kai Zhao, Liang Huang, and Mingbo Ma. 2017. Textual entailment with structured attentions and composition. *arXiv preprint arXiv:1701.01126*.