Can Knowledge Rich Sentences Help Language Models To Solve Common Sense Reasoning Problems?

by

Ashok Prakash

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved October 2019 by the
Graduate Supervisory Committee:

Chitta Baral, Chair
Murthy Devarakonda
Saadat Anwar

ARIZONA STATE UNIVERSITY

December 2019

ABSTRACT

Significance of real-world knowledge for Natural Language Understanding(NLU) is well-known for decades. With advancements in technology, challenging tasks like question-answering, text-summarizing, and machine translation are made possible with continuous efforts in the field of Natural Language Processing(NLP). Yet, knowledge integration to answer common sense questions is still a daunting task. Logical reasoning has been a resort for many of the problems in NLP and has achieved considerable results in the field, but it is difficult to resolve the ambiguities in a natural language. Co-reference resolution is one of the problems where ambiguity arises due to the semantics of the sentence. Another such problem is the cause and result statements which require causal commonsense reasoning to resolve the ambiguity. Modeling these type of problems is not a simple task with rules or logic. State-of-the-art systems addressing these problems use a trained neural network model, which claims to have overall knowledge from a huge trained corpus. These systems answer the questions by using the knowledge embedded in their trained language model. Although the language models embed the knowledge from the data, they use occurrences of words and frequency of co-existing words to solve the prevailing ambiguity. This limits the performance of language models to solve the problems in common-sense reasoning task as it generalizes the concept rather than trying to answer the problem specific to its context. For example, "The painting in Mark's living room shows an oak tree. It is to the right of a house", is a co-reference resolution problem which requires knowledge. Language models can resolve whether "it" refers to "painting" or "tree", since "house" and "tree" are two common co-occurring words so the models can resolve "tree" to be the co-reference. On the other hand, "The large ball crashed right through the table. Because it was made of Styrofoam ." to resolve for "it" which can be either "table" or “ball”, is difficult for a language model as it requires more information about the problem.

In this work, I have built a end-to-end framework, which uses the automatically extracted knowledge based on the problem. This knowledge is augmented with the language models using an explicit reasoning module to resolve the ambiguity. This system is built to improve the accuracy of the language models based approaches for commonsense reasoning. This system has proved to achieve the state of the art accuracy on the Winograd Schema Challenge.

DEDICATION

I dedicate the work to my mom, dad, teachers and my well wishers who have seen me grow this far and supported me when I required the most.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Chitta Baral for his immense support and guidance throughout my master's degree here at ASU. This work has only been possible because of the knowledge I have gained undertaking his course and his guidance. The lessons learnt would help me immensely in my future endeavors. I would also like to thank my committee members Dr. Murthy Devarakonda and Dr. Saadat Anwar, who offered guidance and support whenever I needed. The valuable suggestions and critical comments have helped me to improve this work significantly. I would also like to thank Arpit Sharma, my friend and collaborator, for his constant motivation and answering all my queries patiently throughout this thesis work. I would also like to thank Arindham Mitra for all the encouragement and his guidance. Last but not the least, I would like to thank my mom, dad and sister Vanitha for encouraging me and standing with me through my struggles- words don't do justice to everything you've done for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

In the field of Artificial Intelligence, to given an appropriate response, a set of background information is expected to be known. This background is termed to be commonsense knowledge, this can be shared information by an individual, a group, culture, age, gender, species or substances. To obtain a commonsense knowledge we human beings either learn it or experience it through our day to day activities. For example, when a mad dog chases us, running away from the dog is a natural behavior, because we know that we could get hurt if we stood there. In simple words, it is what people don't have to say because it is expected to know or presumed from experiences.

Currently there have been numerous works on problems which requires commonsense knowledge to solve them (Marcus and Davis 2019). Such problems are an attempt to create a database for several actions, behaviors and responses which is accessible by artificial intelligence programs (Davis and Marcus 2015b) that uses natural language. This task is divided into sub components since the commonsense knowledge is broad and the scope is huge, to replicate the same as a human mind would manage it.

1.1 Motivation

As the problem boils down to understanding natural language, it's the task of the natural language understanding(NLU) to interpret and use such some sort of knowledge for an appropriate action. Over the years there have been many advances and challenges proposed to solve this by NLU (Clark et al. 2018; Mihaylov et al. 2018; Mishra et al. 2018), two such challenges are taken in this work.

- Winograd Schema Challenge (WSC) (Levesque, Davis, and Morgenstern 2011), which is made up of pronoun resolution problems of a particular kind. The main part of each WSC problem is a set of sentences containing a pronoun. In addition, two definite noun phrases, called “answer choices” are also given. The answer choices are part of the input set of sentences. The goal is to determine which answer provides the most natural resolution for the pronoun.

- Choice of Plausible Alternatives (Melissa Roemmele1 and Gordon2 2011), an evaluation of commonsense casual reasoning. Each question is composed of a premise and two alternatives, where the task is to select the alternative that more plausibly has a causal relation with the premise. The correct alternative is randomized so that the expected performance of randomly guessing is 50%.

Below is an example problem from the WSC.

Sentences (S1): The fish ate the worm. It was tasty.

Pronoun to resolve: It

Answer Choices: a) fish b) worm

A WSC problem also specifies a “special word” that occurs in the sentences, and an “alternate word.” Replacing the former by the latter changes the resolution of the pronoun. In the example above, the special word is *tasty* and the alternate word is *hungry*.

The resolution of the pronoun is difficult because the commonsense knowledge that is required to perform the resolution is not explicitly present in the input text. The above example requires the commonsense knowledge that *‘something that is eaten may* *be tasty'*. There are attempts (Sharma et al. 2015; Emami, De La Cruz, et al. 2018) to extract such knowledge from text repositories. Those approach finds the sentences which are similar to the sentences in a WSC problem but without the co-reference ambiguity. For example a sentence (which contains knowledge without ambiguity) corresponding to the above WSC problem is *'John ate a tasty apple'*. Such an approach to extract and use sentences which contain evidence for co-reference resolution is termed as Knowledge Hunting (Sharma et al. 2015; Emami, Trischler, et al. 2018). There are two main modules in the knowledge hunting approach, namely a knowledge extraction module and a reasoning module. To be able to use the extracted knowledge, the reasoning module puts several restrictions on the structure of the knowledge. If the knowledge extraction module could not find any knowledge pertaining to those restrictions, the extracted knowledge would probably be of no use.

Sometimes the needed knowledge are embedded in the pre-trained language models. Let us consider the WSC example mentioned below.

S2: The painting in Mark's living room shows an oak tree. It is to the right of a house.

Pronoun to resolve: It

Answer Choices: a) painting b) tree

Here, the knowledge that *'a tree is to the right of a house'* is more likely than *'a painting is to the right of a house'* is needed. With recent developments in neural network architectures for language modeling, it is evident that they are able to capture such knowledge by predicting that *'a tree is to the right of a house'* is a more probable phrase than *'a painting is to the right of a house'*. This is because language models are trained on huge amounts of text and they are able to learn the frequently co-occurring concepts from that text. Although the knowledge from language models is helpful in many examples, it is not suitable for several others. For example, the language models in (Trinh and Le 2018) predict that *'fish is tasty'* is a more probable than *'worm is tasty'*. This is because the words *'fish'* and *'tasty'* occur in the same context more often than the words *'worm'* and *'tasty'*.

So, considering the benefits and limitations of the above mentioned approaches, in this work, we combine the knowledge hunting and neural language models to solve the Winograd Schema Challenge (WSC). The main contribution of this work is to tackle the WSC by:

Developing and utilizing an automated knowledge hunting approach to extract the needed knowledge and reason with it without relying on a strict formal representation, utilizing the knowledge that is embedded in the language models, and combining the knowledge extracted from knowledge hunting and the knowledge in language models

As a result our approach improves on the existing state of the art accuracy by 7.36% and solves 71.06% of the WSC problems correctly.

1.2 Contributions

- We built a module which could automatically extract knowledge based upon the problem statement and create knowledge corpus for the list of ambiguous problem statements.

- Perfected the existing system suitable to combine with the language model and found which knowledge would work the best for the problem statement.

- Built a system based on Probabilistic Soft Logic(PSL), which provides a confidence score to remove the ambiguity in the problem statement. We came up with the novel approach to combine these alignment pairs produced from the previous system and language models scores.

- Experiments on multiple data set with three different systems are conducted.

- Achieved state of the art accuracy on WSC dataset with this end to end system.

- Analysis on PSL to provide proofs that specific knowledge helps answering the questions. Also, made the analysis and code available public, so the techniques can be applied to similar commonsense reasoning problems.

Chapter 2

BACKGROUND

In past decades computation linguistics have grown to understanding of components of a language. Studies involving Natural Language Understanding(NLU) goes to understanding in the structure of the sentences, phrases, words, discourses. Systems were developed by computational linguists to process these in their natural language form. There are multi-level natural language processing systems the above-mentioned components are handled separately and eventually formed full-proof systems for NLP tasks.

From the basis of these applications, syntactic study of a natural language has been in use for a longer time than using semantics of the language. Linguists who emphasized on the semantic nature of the natural language, believed in resolving multiple problems. Problems in understanding a sentence lies with understanding the semantics of the sentence (Johnson-Laird and Miller 1976). Their firm belief on semantic nature of a natural language, started with figuring out the existing knowledge in the given text.

A NLU system which tries to understand a language uses considerable knowledge about the general world, about the context of the discourse is held in and about the language itself (Allen 1995). (Johnson-Laird and Miller 1976) describes the following in order to explain the need of context in a language:

"Efforts to put some sensible construction on what another person is saying are usually aided by knowledge of the context in which he/she says it. The context provides a pool of shared information on which both parties to a conversation can draw. The

information, both contextual and general, that a speaker believes his listener shares with him constitutes the cognitive background of this utterance."

2.1 Categorization of Knowledge

Understanding of natural language requires various kinds of knowledge. To put it in broader terms there are factual knowledge and commonsense knowledge(Mishra et al. 2018).

2.1.1 Factual Knowledge

A knowledge which is learnt generally over years by seeing and studying them. This knowledge is a fact which is known generally, for example "Earth is the third planet from the sun and earth revolves around the sun" and "Russia is the largest country by land area on earth" are facts.

2.1.2 Commonsense Knowledge

Commonsense knowledge is the knowledge about the concepts of the world. Typical way of acquiring and using such knowledge is through data, which requires saving a data in a form of data structure. One form of such data structure could be using a graphical representation of such concepts. In the below example, it is shown how inference is made out of concepts presented on a graph. First, the task of graph completion was taken place for this problem. When a query is asked to resolve this is intuitively cast as an inference problem from a collection of candidate premises to the truth value of the query. The inference was made from the language which was shown in forward to the system through language constructed in a parsed tree. Here in the figure1, the inference tries to explain the claim that no carnivores eat animals is wrong by inferring to the graph which contains the cat ate a mouse.

In real world, creating such graphical interpretations are considered to be a tedious task. Such inferences are difficult to capture as the task can be huge when split in small problem sets. These inferences are learned and they are made into general rules which can be used to query for exact general which is function of action required to match the premise. New age methods follow storing these information has deep learning models which has learned weights of the sentence forms in vector representation(eg. WordNet).

2.2 Information Extraction

Information extraction(IE) in an open domain has been shown to be useful in a number of NLP tasks, such as question answering, relation extraction and information retrieval (Anthony Fader and Etzioni. 2011), (Soderland et al. 2013). Conventionally, open IE systems search a collection of patterns over either the surface form or dependency tree of a sentence. Although a small set of patterns covers most simple sentences (e.g., subject verb object constructions), relevant relations are often spread across clauses or found in a non-canonical form. To speed up the information extraction process the querying system adds some requirements which take information from the query to search for all combination of the concepts found in the sentence. The order in which the search process was made is preserved and the resultant paragraphs are stored for the language corpus usage. Each answer which if parsed for a specific domain

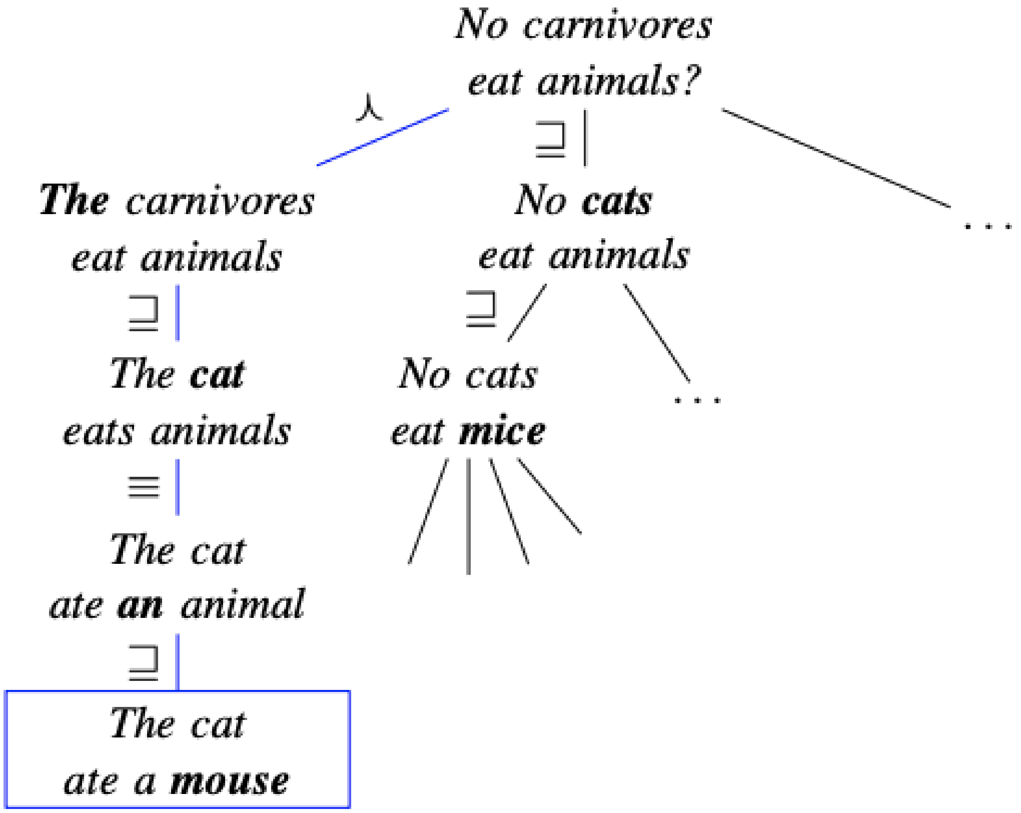


Figure 1. Database containing commonsense for inference *Falsifying the claim that no carnivores eat animals*

is then ranked by their relevancy on the basis of language. Metrics of comparison can be it's similarity in syntactic or semantic representation, entailment, and concept or relation based scoring.

2.3 Natural Language Inference

Natural Language Inference is about figuring out whether the given premise is equal to the hypothesis deducted. This concept is used in entailment and further developments of Language Inference. Specific techniques and developments have been made on this and there are some described as below.

Natural Language Induction has been the focus from the beginning, however while there are programmed techniques for formal methods for improved, there is little advancement at language derivation in Natural Language Inference(NLI). To decide if a characteristic of a language can be legitimately be gathered from a premise. The difficulties of NLI are very not quite the same as those experienced in formal conclusion: the accentuation is on casual thinking, lexical semantic learning, and fluctuation of phonetic articulation.

Investigation have been made in multiple scope of ways to deal with NLI, starting with strategies which are powerful however inexact, and continuing to dynamically increasingly exact methodologies.

Regardless of its outrageous straightforwardness, there are models which accomplishes shockingly great outcomes on a standard NLI assessment, for example, the PASCAL RTE Challenge(Bowman et al. 2015a). In any case, its adequacy is restricted by its inability to speak to semantic structure. To cure this need, Stanford RTE framework(Bowman et al. 2015a) was introduced, which uses composed reliance trees as an intermediary for semantic structure, and looks for a minimal effort arrangement between trees for premise p and hypothesis h, utilizing a cost model which joins both lexical and auxiliary coordinating expenses. One such methodology is to find out entailment between text fragments. The entailing and entailed texts are termed text

(p) and hypothesis (h), respectively. Textual entailment is not the same as pure logical entailment — it has a more relaxed definition: “p entails h” (p -> h) if, typically, a human reading p would infer that h is most likely true. (Alternatively: p -> h if and only if, typically, a human reading p would be justified in inferring the proposition expressed by h from the proposition expressed by t) The relation is directional because even if “p entails h”, the reverse “h entails p” is much less certain.

Determining whether this relationship holds is an informal task, one which sometimes overlaps with the formal tasks of formal semantics (satisfying a strict condition will usually imply satisfaction of a less strict conditioned); additionally, textual entailment partially subsumes word entailment.

An example of a positive TE (text entails hypothesis) is: If you help the needy, God will reward you. hypothesis: Giving money to a poor man has good consequences.

An example of a negative TE (text contradicts hypothesis) is: If you help the needy, God will reward you. hypothesis: Giving money to a poor man has no consequences.

An example of a non-TE (text does not entail nor contradict) is: If you help the needy, God will reward you. hypothesis: Giving money to a poor man will make you a better person.

2.4 Commonsense Reasoning

Significance of acquiring real-world knowledge for natural language processing is discussed since 1960 by Bar-Hillel. In the context of machine translation, commonsense is key for disambiguation of all kinds. As we know that ambiguity in a language exists and it’s interpretation changes the whole context of it’s grounding, need of commonsense urges. A well-known example from Terry Winograd is the pair of sentences known as the winograd schema has multiple examples. One such example is "The city council refused the demonstrators a permit because they feared violence," vs."... because they advocated violence". To resolve the pronoun here what "they" refers for the given scenario they refers to either of them. Thus it proves that just relying on the language clues given in the sentence is not enough to resolve the ambiguity.

Problem of solving ambiguity in the language can span to much larger problems like machine translation, question answering and text summarizing. If commonsense reasoning is used for machine translation, it can provide better systems than directly converting the sentence with their syntax or semantics from the language. (Davis and Marcus 2015a) Google translate tried translating the following sentences "The electrician is working" and "the telephone is working", it's translates "working" into "laboring" in the earlier case and "functioning" in the later. When the sentence "the electrician who came to fix the telephone is working" is translated the same holds true, "working" is converted to "functioning" as "telephone" and "functioning" are most frequent words together. This shows that use of commonsense is vital for language understanding and translation

2.5 Commonsense Causal Reasoning

Theoretical investigations of causality have been pursued across many fields, each helping to refine a definition of causality that agrees with our commonsense intuitions. (Melissa Roemmele1 and Gordon2 2011). In philosophy, a rigorous test for determining a causal relation between two events is that of "necessity in the circumstances". According to this criterion, event A is necessary for eventB if the following statement is true: if A had not occurred in the circumstances, then B would not have occurred (therefore, A causes B). An alternative view of causality requires "sufficiency in the circumstances" between two events. A is said to be sufficient in the circumstances for B if it is true that if A occurs and things continue normally from there, event B will occur (therefore, A causes B). Necessity and sufficiency do seem to play a role in human reasoning about causality, as demonstrated in experimental settings. When subjects detect a relation between two events in terms of necessity and/or sufficiency, they also deem these events as causally related (Thompson, 1989; Trabasso et al., 1989).

2.6 Action and Cause

Another area of commonsense reasoning that is well understood is the theory of action, events, and change. In particular, there are very well established representational and reasoning techniques for domains that satisfy the following constraints:

1. Events are atomic. That is, one event occurs at a time, and the reasoner need only consider the state of the world at the beginning and the end of the event, not the intermediate states while the event is in progress.
2. Every change in the world is the result of an event.
3. Events are deterministic; that is, the state of the world at the end of the event is fully determined by the state of the world at the beginning plus the specification of the event.
4. Single actor. There is only a single actor, and the only events are either his actions or exogenous events in the external environment.

5. Perfect knowledge. The entire relevant state of the world at the start, and all exogenous events are known or can be calculated.

2.7 Next Sentence Prediction

Next Sentence Prediction (NSP) NSP is a classification task to predict if the sentences follow each other. This task is considered to be binary classification loss for predicting whether two segments follow each other in the original text. Alternate examples were created by taking consecutive sentences to check if they follow one another. The text corpus can be used with multiple options or sentence to check if they follow each other or not. On the other case, the negative examples are created by pairing segments from different documents. Both alternate examples which includes positive and negative samples are correlated with equal probability. The NSP objective was designed to improve performance on downstream tasks, such as Natural Language Inference Bowman et al. 2015b, which require reasoning about the relationships between pairs of sentences.

2.8 Probabilistic Models

A statistical language model is a probability distribution over sequences of words. With the sequence of such words we can form distribution of some length. Given such a sequence, say of length m, it assigns a probability to the whole sequence by following equation. The language model provides context to distinguish between words and phrases that sound similar. For example, in American English, the phrases “recognize speech” and “wreck a nice beach” sound similar, but mean different things.

$$P(w_1, \ldots, w_m) \tag{2.1}$$

Data sparsity is a major problem in building language models. Most possible word sequences are not observed in training. One solution is to make the assumption that the probability of a word only depends on the previous n words. This is known as an n-gram model or uni-gram model when n = 1. The uni-gram model is also known as the bag of words model.

Estimating the relative likelihood of different phrases is useful in many natural language processing applications, especially those that generate text as an output. Language modeling is used in speech recognition, machine translation, part-of-speech tagging, parsing, Optical Character Recognition, handwriting recognition, information retrieval and other applications.

In speech recognition, sounds are matched with word sequences. Ambiguities are easier to resolve when evidence from the language model is integrated with a pronunciation model and an acoustic model.

Language models are used in information retrieval in the query likelihood model. There a separate language model is associated with each document in a collection. Documents are ranked based on the probability of the query Q in the document's language model Commonly, the uni-gram language model is used for this purpose.

$$P(Q \mid M_d) \tag{2.2}$$

2.8.1 Probabilistic Soft Logic

PSL is a probabilistic logic framework framework for developing probabilistic models. PSL models are easy to use and fast. The models can be defined using a straightforward logical syntax and solved with fast convex optimization. A key distinguishing feature of PSL is that ground atoms have soft, continuous truth values in the interval [0, 1] rather than binary truth values as used in Markov Logic Networks and most other kinds of probabilistic logic. Given a set of weighted logical formulas, PSL builds a graphical model defining a probability distribution over the continuous space of values of the random variables in the model.

A PSL model is defined using a set of weighted if-then rules in first-order logic, as in the following example:

$$\begin{aligned} 0.7 : \forall x, y, z.spouse(x, y) \wedge isChildOf(z, x) \\ \rightarrow isChildOf(z, y) \end{aligned} \tag{2.3}$$

Here, x, y and z represent variables. The above rule states that a person's child is also a child of his/her spouse. The weight (0.7) associated with the rule encodes the strength of the rule.

Each grounded atom, in a rule of a PSL model has a soft truth value in the interval [0, 1], which is denoted by I(a). Following formulas are used to compute soft truth values for the conjunctions ($\wedge$), disjunctions ($\vee$) and negations ($\neg$) in the logical formulas.

$$
\begin{aligned}
I(l1 \wedge l1) &= max\{0, I(l1) + I(l2) - 1\} \\
I(l1 \vee l1) &= min\{I(l1) + I(l2), 1\} \\
I(\neg l1) &= 1 - I(l1)
\end{aligned}
\tag{2.4}
$$

Then, a given rule r $\equiv$ *rbody* $\rightarrow$ *rhead*, it is said to be satisfied (i.e. I(r) = 1) iff I(*rbody*) $\leq$ I(*rhead*). Otherwise, PSL defines a distance to satisfaction d(r) which captures how far a rule r is from being satisfied: d(r) = max{0, I(*rbody*) - I(*rhead*)}. For example, assume we have the set of evidence: I($spouse(B, A)$) = 1, I($isChildOf(P, B)$) = 0.9, I($isChildOf(P, A)$) = 0.7, and that r is the resulting ground instance of rule (1). Then I($spouse(B, A) \wedge isChildOf(P, B)$)=max{0,1+0.9-1}=0.9, and d(r)=max{0,0.9-0.6}=0.3

2.9 Neural Language Models

Neural language models (or continuous space language models) use continuous representations or embedding of words to make their predictions. These models make use of Neural networks.

Continuous space embedding help to alleviate the curse of dimensionality in language modeling: as language models are trained on larger and larger texts, the number of unique words (the vocabulary) increases.[a] The number of possible sequences of words increases exponentially with the size of the vocabulary, causing a data sparsity problem because of the exponentially many sequences. Thus, statistics are needed to properly estimate probabilities. Neural networks avoid this problem by representing words in a distributed way, as non-linear combinations of weights in a neural net. An alternate description is that a neural net approximates the language function.

The neural net architecture might be feed-forward or recurrent, and while the former is simpler the latter is more common. Typically, neural net language models are constructed and trained as probabilistic classifiers that learn to predict a probability distribution

$$P(w_t|\text{context})\,\forall t \in V \tag{2.5}$$

I.e., the network is trained to predict a probability distribution over the vocabulary, given some linguistic context. This is done using standard neural net training algorithms such as stochastic gradient descent with back-propagation. The context might be a fixed-size window of previous words, so that the network predicts

$$P(w_t|w_{t-k},\ldots,w_{t-1}) \tag{2.6}$$

From a feature vector representing the previous k words. Another option is to use "future" words as well as "past" words as features, so that the estimated probability is 2.7

$$P(w_t|w_{t-k},\ldots,w_{t-1},w_{t+1},\ldots,w_{t+k}) \tag{2.7}$$

A third option that allows faster training is to invert the previous problem and make a neural network learn the context, given a word. One then maximizes the log-probability 2.8.

$$\sum_{-k\leq j-1,\, j\leq k} \log P(w_{t+j}|w_t) \tag{2.8}$$

2.9.1 Bi-Directional Encoder Representations

Bi-Directional Encoder Representations from Transformers(BERT) is a simple, yet powerful language model. It has obtained state-of-the-art results in multiple NLP tasks. In challenging Question Answering problems BERT has proven to work really well than it's peer language models. ((Devlin et al. 2018)) Authors of this paper introduces us to a new method of language model called the BERT and its detailed implementation in this section.

BERT's model architecture(Devlin et al. 2018) is a multi-layer bidirectional Transformer encoder based on the (Vaswani et al. 2017) original implementation described in (Vaswani et al. 2017) and released in the recent tensor flow library. Use of Transformer has become more common nowadays, here the model architecture is explained in details with their number of layers, parameteres used, and also particular model used in this works are highlighted. Model architecture and more details can be found in the (Vaswani et al. 2017)

In this work, we are using the BERT language which is trained on the following architecture with the number of layers (i.e., Transformer blocks) as L, the hidden size as H, and the number of self-attention heads as A. In all cases it is set the feed-forward/filter size to be 4H, i.e., 3072 for the H = 768 and 4096 for the H = 1024. We primarily report results on two model sizes:

- BASE: L=12, H=768, A=12, Total Parameters=110M
- LARGE: L=24, H=1024, A=16, Total Parameters=340M

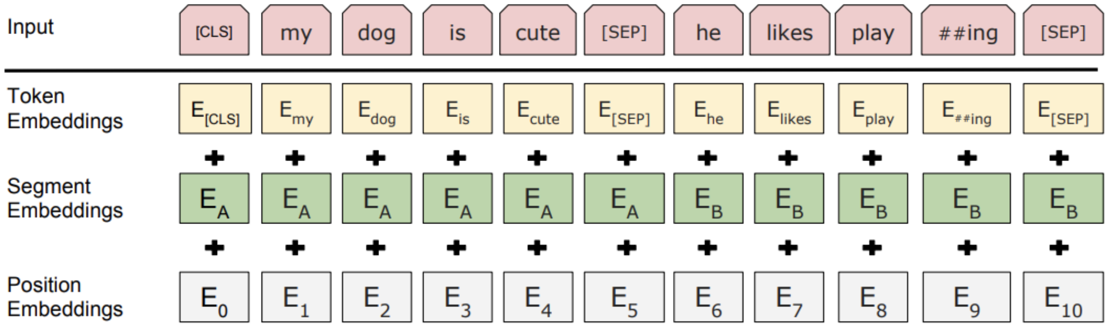


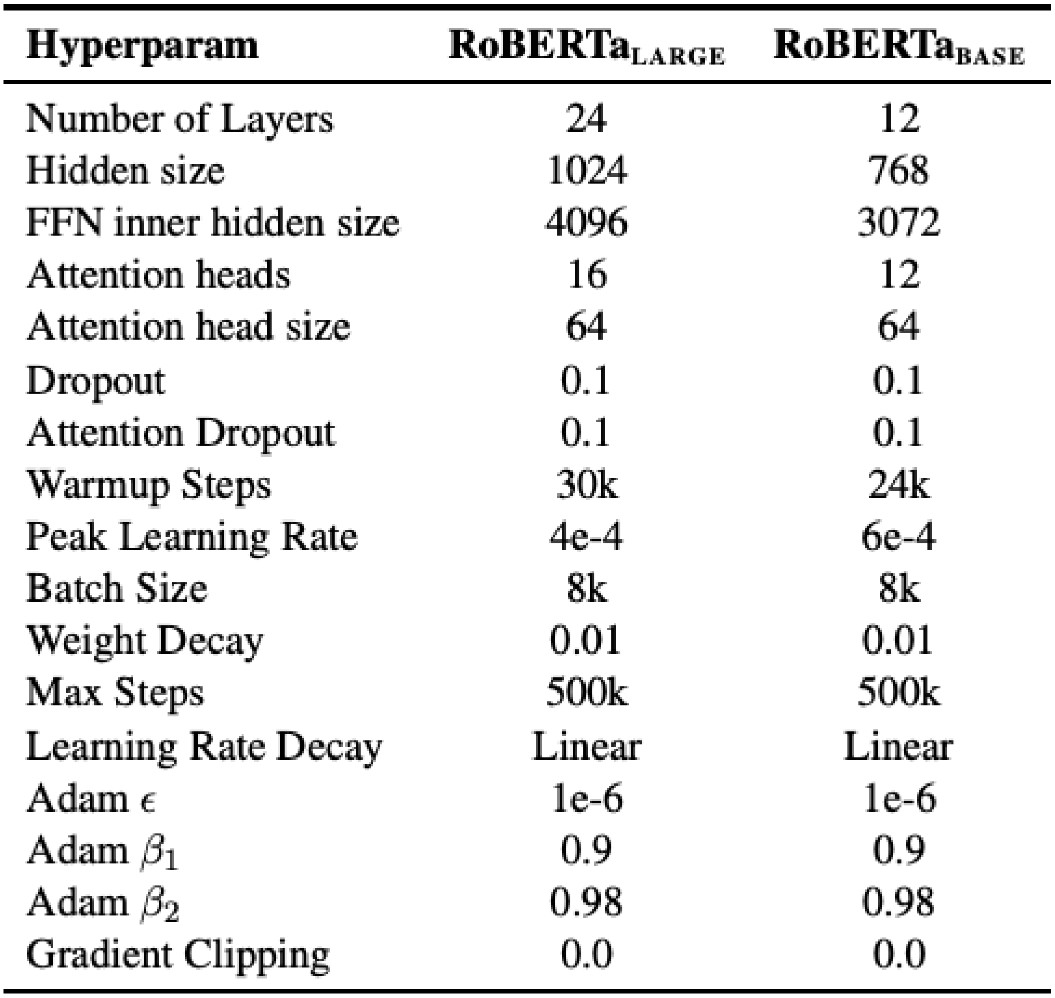
Figure 2. Input representation in BERT models

2.9.2 RoBERTa

RoBERTa is an improved version of the BERT style pretrained models, which tries the improve the model accuracy through adding more data and also optimizing it work much faster. Y. Liu et al. 2019 The works of the authors mostly relies on training of larger datasets where Roberta acts the same way as BERT in abstract terms. The authors have considered fuve english-language corpora for this purpose and the collected 160GB of text to be trained on this model. This four times larger than the actual BERT-Large model which we discussed earlier. Roberta as we speak, is going through multiple iterations and advancements towards adding more data and improvising the model. Adam optimizations has been done on the same to earn maximum potential in the model. The following copora is used for the training purpose, which are as follows.

Below network architecture for RoBERTa is shown from (Y. Liu et al. 2019)

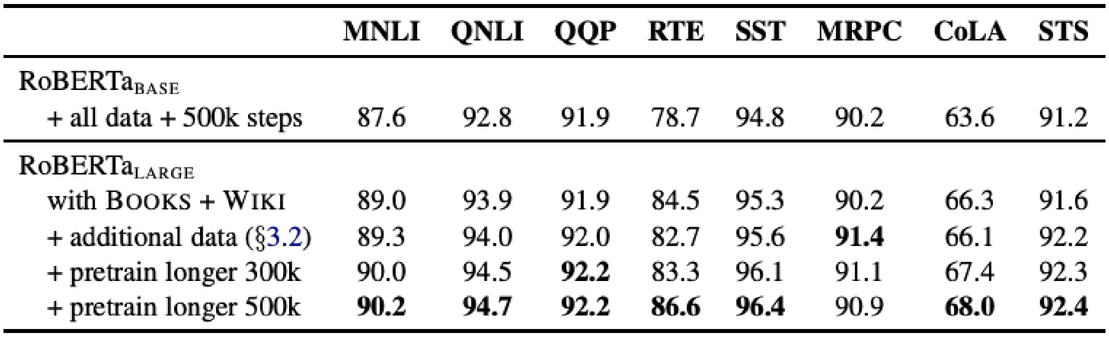
1. BOOKCORPUS (Zhu et al. 2015) is almost 16GB in size which is used in the BERT training phase, along with Wikipedia



Hyperparam	RoBERTa$_{\text{LARGE}}$	RoBERTa$_{\text{BASE}}$
Number of Layers	24	12
Hidden size	1024	768
FFN inner hidden size	4096	3072
Attention heads	16	12
Attention head size	64	64
Dropout	0.1	0.1
Attention Dropout	0.1	0.1
Warmup Steps	30k	24k
Peak Learning Rate	4e-4	6e-4
Batch Size	8k	8k
Weight Decay	0.01	0.01
Max Steps	500k	500k
Learning Rate Decay	Linear	Linear
Adam ϵ	1e-6	1e-6
Adam β_1	0.9	0.9
Adam β_2	0.98	0.98
Gradient Clipping	0.0	0.0

Figure 3. RoBERTa network architecture

2. CommonCrawl News was used, which contained 63 million news articles between September 2016 and February 2019
3. OPENWEBTEXT (Gokaslan and Cohen 2019) which is a text corpus containing the texts from forum with minimum three upvotes. The total size of this dataset is 38GB



	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS
RoBERTa$_{\text{BASE}}$								
+ all data + 500k steps	87.6	92.8	91.9	78.7	94.8	90.2	63.6	91.2
RoBERTa$_{\text{LARGE}}$								
with BOOKS + WIKI	89.0	93.9	91.9	84.5	95.3	90.2	66.3	91.6
+ additional data (§3.2)	89.3	94.0	92.0	82.7	95.6	**91.4**	66.1	92.2
+ pretrain longer 300k	90.0	94.5	**92.2**	83.3	96.1	91.1	67.4	92.3
+ pretrain longer 500k	**90.2**	**94.7**	**92.2**	**86.6**	**96.4**	90.9	**68.0**	**92.4**

Figure 4. RoBERTa linear tasks performed on the above datasets

4. STORIES contained Winograd like stories and information filtered from CommnCrawl

Chapter 3

RELATED AND EXISTING METHODS

Continuous efforts in the field of Natural Language Processing has given tremendous improvement to the field. In this chapter, I look forward to explain the related works conducted in commonsense reasoning, language models and knowledge extraction.

3.1 Knowledge Extraction For Commonsense Reasoning

Some works on knowledge extraction has led to the path of figuring commonsense knowledge from huge text corpus. For instance, databases has been created from the Wikipedia text corpus and several data mining techniques have been employed to extract knowledge from the text. As in the works of (Robert Speer and Lieberman 2008) aims to reduce the noise of the text extracted and tries to use it subjective. This is achieved by forming analogical closure of a semantic network through dimensionality reduction. It self-arranges entities around dimensions that can be viewed as making refinements, for example, "good vs bad" or "easy vs hard", and sums up its learning by making a decision about where ideas lie along these dimensions.

Their assessment shows that users frequently concur with the anticipated information. While works to fix the curse of dimensionality since the web is huge has landed them to refine the knowledge and identify concepts, the concepts extracted were still unclear and more efforts were required to make sense on the analogy. The problems of word sense disambiguation prevailed which required construction of a proper knowledge base to the chosen problem in hand. In the efforts of (Singh et al. 2002) has led to open mind common sense which is a knowledge acquisition framework intended to get commonsense knowledge from the overall network of the web. They depict and assess their first handled framework, which empowered the development of a 450,000 declaration conventional information base. They point-out how their second-generation framework tends to improve on the shortcomings found in the first. The new framework secures facts, descriptions, and stories by enabling members to develop and fill in regular language formats. It utilizes word-sense disambiguation and strategies for clearing up entered information, analogical derivation to give input, and enables members to approve learning and thus one another.

3.2 Relational And Logical Models

Information extraction from the knowledge base requires to be inferred. Inferring the right knowledge for the right context is a difficult task. Generalizing this task through a logical template to make the learning of commonsense more effective works of (Roni Khardon and z 1999) has view of learning and reasoning with relation representation in the context of NLP. Their works on NLP combined with Inductive Logic Programming(ILP) explains the relationships on the extracted sentence to the world knowledge theoretically. They also state that relationship learning between concepts can provide scalable approach for better functions. Works on ILP has gone for decades and eventually more works on relational and logical learning took place through Answer Set Programming(ASP). (Gelfond and Kahl 2014). In this it explains the notion of logic through the programming language to represent the world terms as objects, functions and relations. The clear separation of the semantics through the syntax in the language help create atomic properties to the individual functions and generalize it's place when it requires through the knowledge. Modeling the domain with relations, identifying the separation between the open and close domain has led relation learning solve multiple problems in this field.

The Winograd Schema Challenge is an alternative to the Turing Test that may provide a more meaningful measure of machine intelligence. (Daniel Bailey et al. 2015) It poses a set of coreference resolution problems that cannot be solved without human-like reasoning. The authors take the view that the solution to such problems lies in establishing discourse coherence. Specifically to examine two types of rhetorical relations that can be used to establish discourse coherence: positive and negative correlation. They introduce a framework for reasoning about correlation between sentences, and show how this framework can be used to justify solutions to some Winograd Schema problems.

3.3 Neural Language Models

Language models sole purpose is to compute the probability of a particular token occurring in the query. It can be sentence or sequence of words which provides the grammar of the language and finds probability of the word following or occurring whichever place the model intends to solve.

Below is the sample neural network architecture shown from neurallmmodelexplained

Neural language models are a basic piece of numerous frameworks that endeavor to understand characteristic language handling task, for example, machine translation and speech recognition. Right now, all cutting edge language models are neural networks.

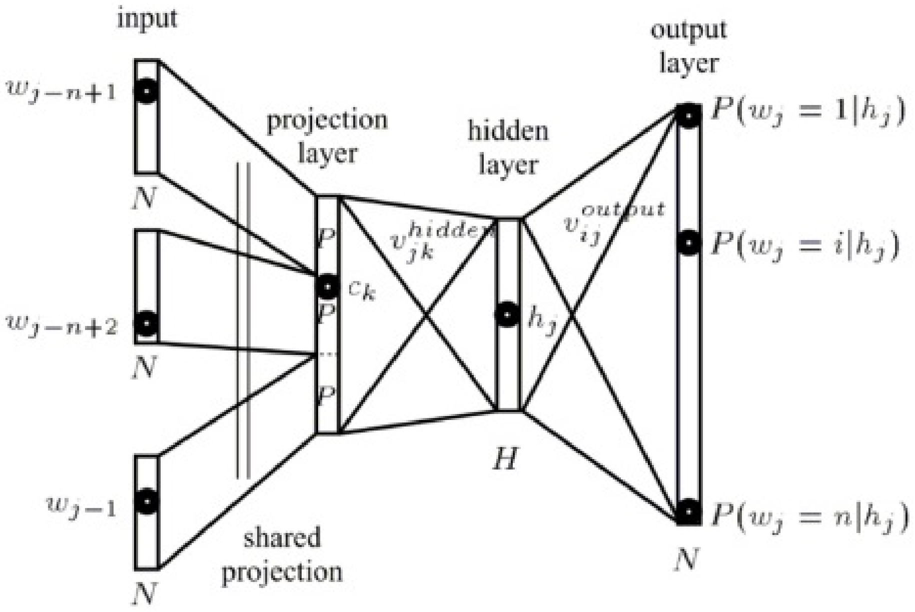


Figure 5. Neural Network Language Model Architecture *Input Layer, Hidden Layer and the Output Layer*

To predict the words in the sentence, neural network take a huge chunks of text has vectors, these vectors can be of any working representations. Commonly used forms are the one-hot vectors. Which does arbitrary ordering of the words in the vocabulary and then represent the nth word as a vector of the size of the vocabulary (N), which is set to 0 everywhere except element n which is set to 1.

Then the model is split in two major components, first component being the encoder and second component is the decoder. To encode the input word, the one hot vector representing the of work of form let's say N X M is taken. This representation is called the input embedding. This multiplication results in a vector of size M, which is also referred to as a word embedding. This embedding is dense representation of the current word given. This representation is both of an a lot littler size than the one-hot vector representing the same word, and furthermore has some other intriguing properties. For instance, while the distance between each two words spoken to by a one-hot vectors is dependably the equivalent, these dense representations have the property that words that are close in significance will have representations that are close in the embedding layer.

After the encoding of the input work in to encoding layer, the embedding which is created in the embedding layer is decoded. The representation of the input word is multiplied by M X N which is the output embedding. The results of the this multiplication is of the size N which passed to a softmax function, this function can be ReLU, sigmoid or any logical function which normalizes the values of the multiplied factors between 0 and 1, and their sum is also equal to 1. This is show in the figure 2 from the article.

Basic function of the decoder is to take a representation of the input word and returns a distribution which speaks to the model's predictions for the following word: the model allocates to each word the likelihood that it will be the following word in the arrangement.

To train a language model, we usually need a set of information and target outputs which are required to be predicted. For Instance, a dataset with the (input, expected output) is taken which at least contains some basic sentences to provide the vocabulary required by the language model. To create word sets for the model to gain from, each pair of the neighbouring words or windows of the sentences are taken has multiple batches and fed into the network has mentioned in the above procedure. Once the information is provided for every input sentence or pair X the expected word Y is provided. Let's say "Alice went to the jungle with an horse" with bi-gram sets the information would like this (Alice, went), (went, to), (to, the),and so on. Once the input is embedded inside the model, stochastic gradient descent is used to update the model in the process of the training. The loss values are noted for every iteration of the process and continued till the loss function yields the minimal value. The loss is measured between the output of the model from the expected output given in the starting phase.

For reporting the performance of the language model, the typical metric is to calculate the perplexity of the model on the test set. It is the likelihood given by the model to the ith target word. Perplexity is a diminishing function of the normal log likelihood that the model relegates to each objective word. We need to expand the likelihood that we provide for each objective word, which implies that we need to limit the perplexity.

Commonsense reasoning is a challenging task which is took it's turn from logical reasoning to deep learning. Recently almost every method which has the state-of-the-art accuracy is a neural network model. Even though the models exist for commonsense reasoning solving problem like winograd schema challenge(Levesque, Davis, and Morgenstern 2011) is difficult. There are works on this particular problem to better use the commonsense and solve ambiguities in the language. Following are some of the works done on these problems.

3.4 Unsupervised Multitask Learners

In this paper, (Trinh and Le 2018) the authors present a simple method for commonsense reasoning with neural networks, using unsupervised learning. Key to our method is the use of language models, trained on a massive amount of unlabled data, to score multiple choice questions posed by commonsense reasoning tests. On both Pronoun Disambiguation and Winograd Schema challenges, our models outperform previous state-of-the-art methods by a large margin, without using expensive annotated knowledge bases or hand-engineered features. Large array of RNN language models are trained to operate at word or character level on LM-1-Billion, CommonCrawl, SQuAD, Gutenberg Books, and a customized corpus for this task and show that diversity of training data plays an important role in test performance. More analysis on the language model shows that the system successfully discovers important features of the context that decide the correct answer, indicating a good grasp of commonsense knowledge.

(Radford et al. 2019) Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task specific datasets. Language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. The largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText. Samples from the model reflect these improvements and contain coherent paragraphs of text. These findings suggest a promising path towards building language processing systems which learn to perform tasks from their naturally occurring demonstrations.

Chapter 4

PROPOSED METHOD

4.1 An Overview On System Architecture

In this work, the target is to build an end to end system which can use the commonsense knowledge more effectively to resolve the ambiguity. In figure 3, the system architecture is shown, which describes the initial problem statement fed into the system. The problems goes to both the specific knowledge system as well as the language models to combine the scores later by the PSL system to provide the confidence. In the specific knowledge model, its presented with few components which helps understand the knowledge and interpret between good knowledge which can help resolve the question and the bad knowledge which override the answer from a language or not solve the ambiguity.

To achieve the above, knowledge for the specific problem is extracted and passed onto the alignment generation model. Here with the use of QASRL, entailment and textual similarities an alignment pair is created between the problem and the knowledge extracted. This alignment pair is saved for every knowledge statement extracted to verify whether the knowledge extracted could help resolve the problem statement. This particular process is later well explained with the WSC problems in the later chapter. With these alignment pairs as output from the specific knowledge system and the language models probability scores for the options the PSL determines the final confidence of the solution.

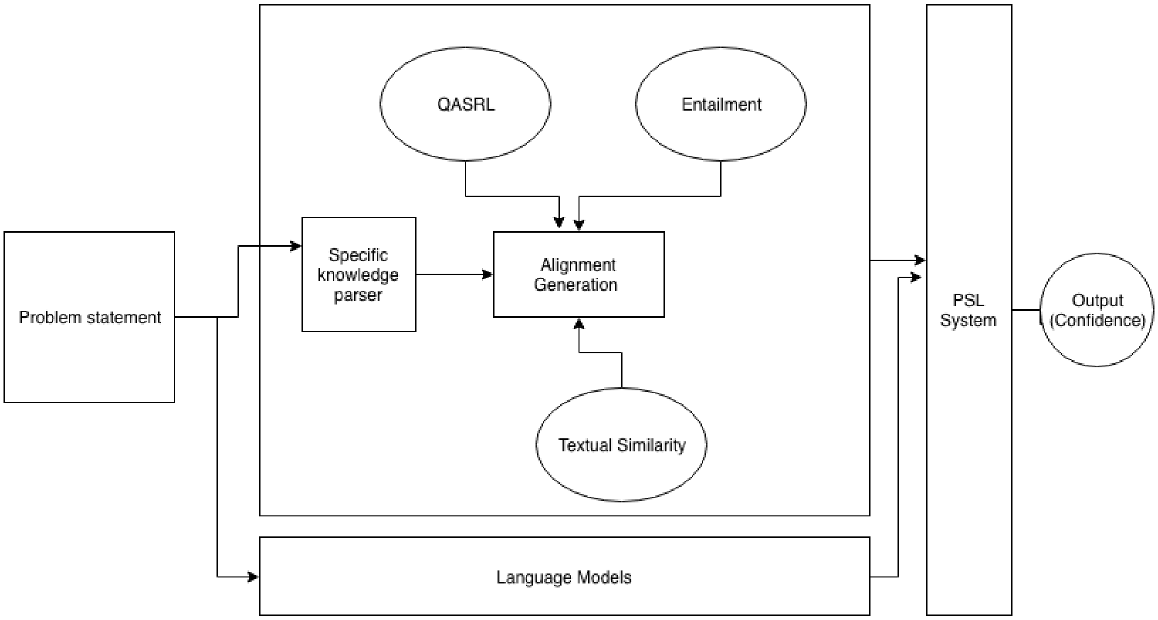


Figure 6. Overall system for solving ambiguity in language

4.2 Dataset

4.2.1 Winograd Schema Challenge

The Winograd Schema Challenge is a co-reference resolution problem. The problem of co-reference resolution has received large amount of attention in the field of Natural Language Processing (Raghunathan et al. 2010; Carbonell and Brown 1988; Ng 2017). However the requirement to use commonsense knowledge makes the Winograd Schema Challenge hard and the other approaches that are trained on their respective corpora do not perform well in the Winograd Schema problems.

The Winograd Schema Challenge was first proposed in 2011 and since then various works have been proposed to address it. These approaches can be broadly categorized into two types:

- The approaches which use explicit commonsense knowledge and reasoning with the knowledge. Such approaches can further be divided into two types.

(a) The approaches which provide a reasoning theory (Dan Bailey et al. 2015; Schüller 2014; Sharma et al. 2015) with respect to a few specific types of commonsense knowledge and takes question specific knowledge while solving a Winograd Schema problem. One of the major shortcomings of such approaches is that they work only for the specific knowledge types and hence their coverage is restricted. Another shortcoming of such approaches is that they rely on strict formal representations of natural language text. The automatic development of such representations boils down to the well known complex problem of translating a natural language text into its formal meaning representation. Among these works, only the work of (Sharma et al. 2015) accepts natural language knowledge sentences which it automatically converts into their required representation. The remaining two (Dan Bailey et al. 2015; Schüller 2014) requires the knowledge to be provided in a logical form.

(b) These approaches (Isaak and Michael 2016) also answer a Winograd Schema problem with formal reasoning but use an existing knowledge base of facts and first-order rules to do that.

- These approaches (Q. Liu et al. 2017; Trinh and Le 2018) utilize the recent advancement in the field of neural networks, particularly the benefits of word embedding and neural language model. The work of (Q. Liu et al. 2017) uses ConceptNet and raw texts to train word embeddings which they later use to solve a Winograd Schema problem by a simple inference algorithm. The work of (Trinh and Le 2018) on the other hand uses majority voting from several language models to resolve the co-reference. In layman terms, the system in (Trinh and Le 2018) replaces the pronoun with the two answer choices to obtain two different sentences and then use the language models to find out which of the two replacement is more probable.

4.2.2 Choice Of Plausible Alternatives

The Choice of Plausible Alternatives (COPA) evaluation Gordon, Kozareva, and Roemmele 2012 consists of 1000 questions of commonsense causality question with two alternative to be chosen. The question set was made utilizing a particular writing a strategy that guaranteed broadness of aspects, lucidity of the language, and high understanding among human raters. This segment portrays the creating approach, concentrating on issues of broadness, clearness, and understanding.

Gordon, Kozareva, and Roemmele 2012 The examples below are taken from the works mentioned above.

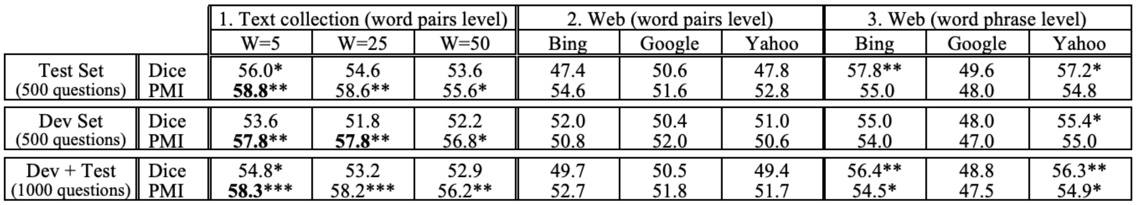
(forward causal reasoning) Premise: The man lost his balance on the ladder. What happened as a result? Alternative 1: He fell off the ladder. Alternative 2: He climbed up the ladder.

(backwards causal reasoning) Premise: The man fell unconscious. What was the cause of this? Alternative 1: The assailant struck the man in the head. Alternative 2: The assailant took the man's wallet.

The primary works of the authors focuses on creating a system was the expansiveness of the question set. The methodology was to distinguish question themes from various sources where a high level of abstractness was brought into act, and after that expand these themes into premises and options through our very own imagination. The author's approach helped balance the logical and generative parts of this errand, guaranteeing that the specific subject interests of the creator were not over-spoke to in the inquiry set, yet at the same time taking into consideration the plan arrangements that every one of these inquiries required. Two essential primary sources of question points were utilized to guarantee breadth.

First of the authors took multiple randomized topics available on the Internet weblogs. To include the social, mental, physical and natural causality as much as possible to make the dataset diverse.

Below tables show the comparison for COPA evaluation from (Melissa Roemmele1 and Gordon2 2011)



		1. Text collection (word pairs level)			2. Web (word pairs level)			3. Web (word phrase level)		
		W=5	W=25	W=50	Bing	Google	Yahoo	Bing	Google	Yahoo
Test Set (500 questions)	Dice	56.0*	54.6	53.6	47.4	50.6	47.8	57.8**	49.6	57.2*
	PMI	58.8**	58.6**	55.6*	54.6	51.6	52.8	55.0	48.0	54.8
Dev Set (500 questions)	Dice	53.6	51.8	52.2	52.0	50.4	51.0	55.0	48.0	55.4*
	PMI	57.8**	57.8**	56.8*	50.8	52.0	50.6	54.0	47.0	55.0
Dev + Test (1000 questions)	Dice	54.8*	53.2	52.9	49.7	50.5	49.4	56.4**	48.8	56.3**
	PMI	58.3***	58.2***	56.2**	52.7	51.8	51.7	54.5*	47.5	54.9*

Figure 7. Three baseline results for the COPA evaluation. The results are computed in terms of accuracy and the ones marked with ***, **, and * are statistically significantly better at the 0.001, 0.01, and 0.05 levels, respectively, than the random baseline (50% accuracy))

From each of these question topics, a pair of statements (the premise and the correct alternative) that captured a key causal relationship. This part of the task required subjective creativity, guided by introspective questions about the topic. For instance, from the topic of "unconsciousness" the authors asked themselves "what causes unconsciousness?" and "what does unconsciousness cause?" Melissa Roemmele1 and Gordon2 2011) Answers to these questions were treated as a causal bridging inference, e.g. "injuries to the head cause unconsciousness." From this, a suitable premise and correct alternative could be instantiated as the events of the causal relation, e.g. "the assailant struck the man in the head" and "the man fell unconscious."

Either the cause or the effect in this pair could be treated as the premise, depending on whether the question was testing forward or backward causal reasoning

The characteristic language expression of every one of the questions information had various rules and number of constructive guidelines to include expansions in the dataset. The reason and the choices were written in the past tense. They were as brief as could be allowed, overlooking words that were not important to choose the right option. Appropriate names of individuals and spots were maintained a strategic distance from, as were expressions and slang. Individual pronouns and unequivocal determiners were utilized, which drove us to embrace a specific style for co-reference and anaphora.

For example, consider the following question: Premise: The man dropped food on the floor. What happened as a result? Alternative 1. His dog ran over to eat the food. Alternative 2. His dog jumped up on him

The alternatives for this question both explicitly reference a dog whose existence must be presumed in the premise. Here the personal and possessive pronouns ("his", "him") must be resolved to "the man", and "the food" must be seen as co-referential with "food" in the premise.

4.2.3 Winogrande: An Adversarial Winograd Schema Challenge Dataset

WINOGRANDE Sakaguchi et al. 2019is a newly collected dataset which is very similar to the WSC problems that are formed to be robust against the biases such as language based and dataset specific bias. This dataset when compared to the original Winograd dataset and the variants of the same it presents more challenging and more importantly it has a large number of problems through the crowdsourcing.

This particular dataset is also chosen for the study and analysis after the work of Prakash et al. 2019, because in the previous work the WSC datset is studied and the experiments are also discussed as follows. Yet adversarial dataset like WINOGRANDE can prove how better the system work when such examples are introduced. Moreover the improvement of the dataset from 273 to 44K examples is a great number to include a lot of real world scenario for putting the method into test.

4.3 Automated Knowledge Extraction

In this section we first explain how our knowledge hunting approach and the neural language models are used to generate an intermediate results. Then the details of a Probabilistic Soft Logic module which combines the intermediate results and predicts the confidence for each of the answer choices in a WSC example are explained. Moreover how similiar methods can be employed in Causal Reasoning is also demonstrated in the following.

4.3.1 Automated Approach

There are two main modules in the Knowledge Hunting approach. The first module extracts a set of sentences corresponding to a problem sentence such that the extracted sentences may contain the needed commonsense knowledge. We call such a set of sentences, a *knowledge text*. The second module uses a *knowledge text* and generates a correspondence between the answer choices and find contexts to resolve ambiguities in a problem text through identifying the entities in a *knowledge text*. We call such a correspondence as alignment. Such an alignment is an intermediate result from the knowledge hunting module.

1. In the case of WSC sentence, the pronoun is the factor which has ambiguity. To resolve this ambiguity, we keep the pronoun in place to extract the knowledge required.
2. In the COPA dataset, causal reasoning has to be done where cause and effect is given as question or as an alternative. Since there is ambiguity exists in the alternatives provided, we find knowledge sentences for both the alternatives appended with the question.

4.3.1.1 Knowledge Extraction

The goal of the knowledge extraction module is to automatically extract a set of knowledge texts for a given problem sentence. Ideally, a *knowledge text* should be able to justify the answer of the associated problem. In this approach, we aim to extract the texts that depict a scenario that is similar to that of the associated problem sentence which can further resolve the ambiguity in the language. We roughly characterize a problem scenario in terms of the events (verb phrases) and the properties of the entities that are associated with the scenario. The characterization of a scenario optionally includes the discourse connectives between the events and properties of the scenario.

For example, in the WSC sentence *"The city councilmen refused the demonstrators a permit because they feared violence ."*, the scenario is mainly characterized by the verb phrases *"refused"* and *"feared"*, and the discourse connective *"because"*.

In this work, we use this abstract notion of a scenario to extract *knowledge texts* which depict similar scenarios. The following are the steps in the extraction module.

1. First, the module identifies the verb phrases, properties and discourse connectives in a given WSC *scenario*. For example the verb phrases *"refused"* and *"feared"*, and the discourse connective *"because"* in the example mentioned above.
2. Secondly, the module automatically generates a set of search queries by using the keywords extracted in the previous step. The first query in the set is an ordered combination (as per the WSC sentence) of the keywords extracted in the previous step. For example the query *"* refused * because * feared * "* is the first query for the problem mentioned above. Afterwards the following set of modifications are performed with respect to the first query and the results are added to the set of queries. • The verb phrases are converted to their base form. For example, *" * refuse * because * fear * "*.

   • The discourse connectives are omitted. For example, *"* refuse * fear * "*.

   • The verbs in verb phrases and the adjectives are replaced with their synonyms from the WordNet KB (Miller 1995). The top five synonyms from the top synset of the same part of speech are considered. An example query generated after this step is *"* decline * because * fear * "*.
3. Thirdly, the module uses the generated queries to search and extract text snippets, of length up to 30 words, from a search engine. In this step we wanted to filter the top 10 sentences which represent the same or similar scenario to that of the WSC. But we do not want the extracted text to be same as the input WSC text. Neither do we want the extracted text to contain similar co-reference ambiguities as the WSC text. So, we added constrains to filter out the unwanted text. First, So, we made sure that if the extracted text is same as a WSC text

then it is filtered out. Secondly, the extracted text may contain pronouns which may be responsible for a co-reference ambiguity. For example the extracted sentence "I can eat it because it is tasty" also contains two pronoun occurrences (*"it"*,). We make sure that the pronouns in the extracted texts can be easily resolved using the following procedure:

a) Two pronouns refer to each other if they have the same string description. For e.g. all the occurrences of the pronoun "it" always refer to the same entity.

b) Two pronouns (p_1, p_2) refer to each other, if they are of same gender, group, object, which they belong to a special list containing the following: $\{(he, him), (she, her), (i, me), (they, them), (he, his), (his, him)\}$. We also ignore knowledge sentences where any of these special pair of pronouns appears as an argument to a common verb (e.g. "it ate it because ...").

In this work, we have used the search engine of Google.

After several iterations of step 2 and step 3 we obtain sentences which can justify the answer of the given WSCR and do not contain the coreference ambiguity which was present in the original WSCR sentence.

An example *knowledge text* extracted by using the query *" * refused * because * feared * "* via the steps mentioned above is, *"He also refused to give his full name because he feared for his safety."*

Similarly, if the extracted sentence was *"They could not lift it off the floor because they were weak"* then the number agreement (both entities plural or singular) can be used to corefer. We call such texts as *"knowledge texts"*. The manually curated knowledge base contains only *"easily resolvable"* and no-pronoun sentences. Table 1 shows some of the sample knowledge sentences and the corresponding WSC problem.

Problem	Query	Knowledge
Mary tucked her daughter Anne into bed , so that she could sleep	‘*tucked*sleep*’, ‘*tuck*sleep*’	Sleep is essential for physical performance, cognitive functioning, and mental health, tucking yourselves helps you sleep better
Joan made sure to thank Susan for all the help she had received.	‘*thank.*because.*helped.*’	I just wanted to thank you because you have helped me the best when I received the deny letter from immigration some years ago.
Paul tried to call George on the phone , but he was not ’ available.	‘* call * but * unavailable * ’	Every half hour she ran out to call Jack , but his phone was unavailable.
Although they ran at about the same speed , Sue beat Sally because she had such a bad start.	‘* I lost * I had a bad start *’	Granted they were not anchored down and encountered heavy winds, so whether they rolled and then collapsed , or the increased tension ripped the hubs
Frank was upset with Tom because the toaster he had bought from him did not work .	‘ * frustrated * bought * didn’t work *’	I would see people frustrated when they bought products that did not work or super happy when they bought products that made them

Table 1. Example knowledge extracted from the problem sentences

4.3.2 Knowledge Ranking

Since we get numerous such *Knowledge texts* from the obtained search query, we need to rank them for usable knowledge texts. We here compare the knowledge text with the problem sentence in two ways and rank them using the following methods.

1. ELMO entailment(Peters et al. 2017) is used where a premise and hypothesis is used to check the entailment score between the two.

   The model take a pair of sentences and predict whether the facts in the first necessarily imply the facts in the second. Here in this task ELMo (Peters et al. 2017) based language models is used to find the textual entailment with the knowledge and problem statement. (Elmo tackles of problem of polysemy in the Natural Language, which has multiple meanings for the same word. It has complex Bi-Directional LSTM architectures. This architecture can hold multiple embedding for the same word which has different meaning.)

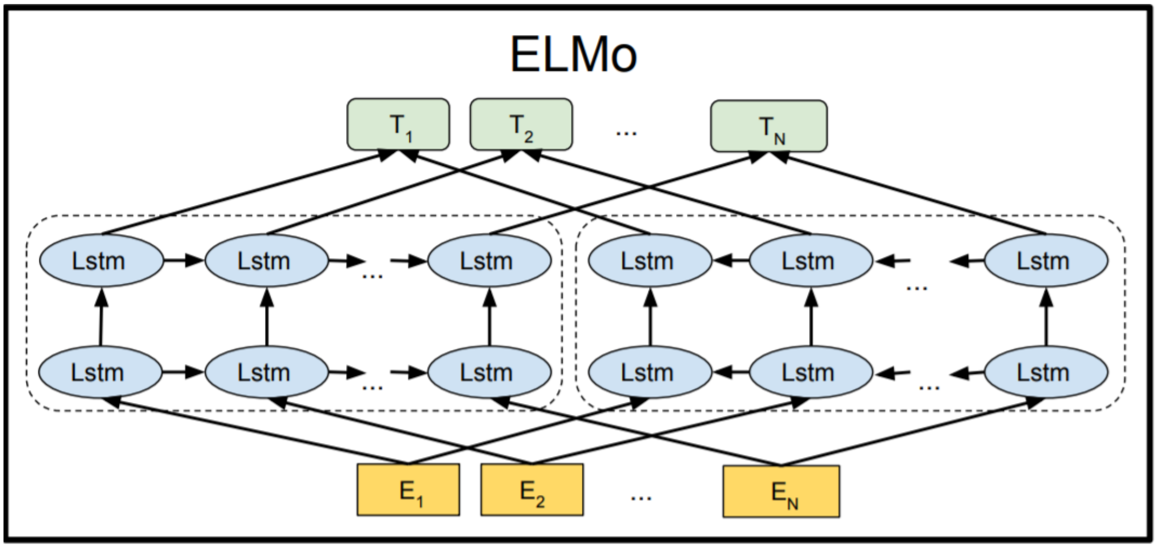


Figure 8. Shallow feed forward network to train for context embedding in ELMO architecture

2. BERT semantic sentence similarity(STS) model is used to check the semantic similarity between the knowledge text obtained and the problem statement is identified. BERT STS is trained on a semantic textual similarity dataset dedicated for this purpose. It gives a score between 0-1 for a sentence being similar to sentence being compared.

Using the scoring method we rank the knowledge and only the top 10 knowledge texts are used in the further process.

1. In the case of ELMO, there are entailment, neutral and contradiction scores. We use the entailment scores for the ranking
2. In the case of BERT STS, we use the similarity score from the model

Sentences (S1): Mary tucked her daughter Anne into bed, so that she could sleep.

Pronoun to resolve: She

Answer Choices: a) Mary b) Anne

For the above problem which is presented from the winograd data set, below are the examples of knowledge texts extracted.

1. *Sleep is essential for physical performance, cognitive functioning, and mental health, tucking yourselves helps you sleep better*
2. *cat seemed very serious about being tucked in to sleep by Ron in the morning and at night*
3. *when I came across an interesting question about people who tuck in the top sheet on their beds*
4. *do You Sleep with Your Top Sheet Tucked or Untucked?*
5. *as a child I remember the comfort I experienced being tucked in at night with great tenderness and loving by my mother. As an adult, I have created a way to tuck myself in with the same sweetness and loving.*
6. even more.....

After using the knowledge ranking method, the same texts which are retrieved are sorted in the order of higher scores to lower. Below it is shown for BERT STS scores.

1. *cat seemed very serious about being tucked in to sleep by Ron in the morning and at night* : 0.80
2. *when I came across an interesting question about people who tuck in the top sheet on their beds*: 0.55
3. *Sleep is essential for physical performance, cognitive functioning, and mental health, tucking yourselves helps you sleep better*: 0.54

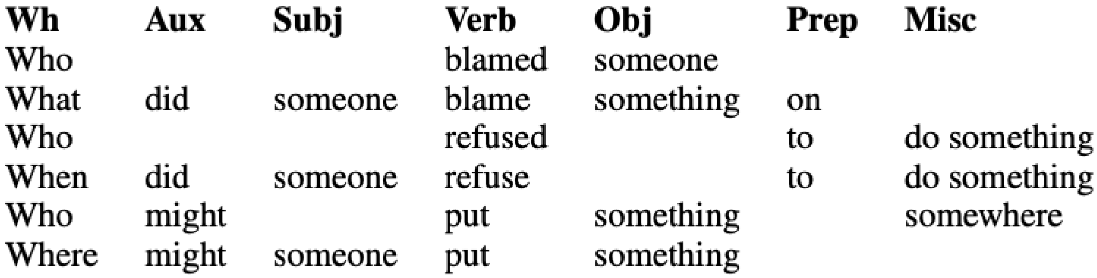
4. *as a child I remember the comfort I experienced being tucked in at night with great tenderness and loving by my mother. As an adult, I have created a way to tuck myself in with the same sweetness and loving.* : 0.49
5. *do You Sleep with Your Top Sheet Tucked or Untucked?* : 0.44
6. even more.....

4.3.3 QASRL

Semantic role labeling (SRL) is the widely studied challenge of recovering predicate-argument structure for natural language words, typically verbs. (He, Lewis, and Zettlemoyer 2015) The goal is to determine "who does what to whom," "when," and "where," etc. This paper introduces the task of questionanswer driven semantic role labeling (QA-SRL), where question-answer pairs are used to represent predicate-argument structure. For example, the verb "introduce" in the previous sentence would be labeled with the questions "What is introduced?", and "What introduces something?", each paired with the phrase from the sentence that gives the correct answer. Posing the problem this way allows the questions themselves to define the set of possible roles, without the need for predefined frame or thematic role ontologies. It also allows for scalable data collection by annotators with very little training and no linguistic expertise. We gather data in two domains, newswire text and Wikipedia articles, and introduce simple classifierbased models for predicting which questions to ask and what their answers should be. Our results show that non-expert annotators can produce high quality QA-SRL data, and also establish baseline performance levels for future work on this task.

Question Answering Semantic Role Labelling, question types with verb tense in *present, past participle and other forms.*

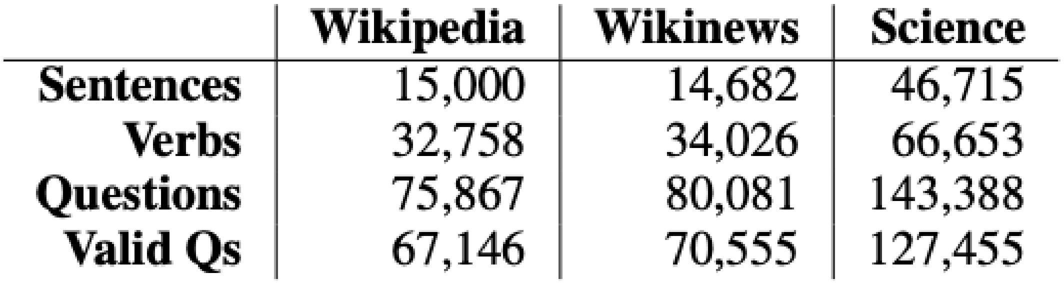
Below example for the QA-SRL for the outcomes are shown from (He, Lewis, and Zettlemoyer 2015)



Wh	**Aux**	**Subj**	**Verb**	**Obj**	**Prep**	**Misc**
Who			blamed	someone		
What	did	someone	blame	something	on	
Who			refused		to	do something
When	did	someone	refuse		to	do something
Who	might		put	something		somewhere
Where	might	someone	put	something		

Figure 9. Example questions in QA-SRL with their possible outcomes in the given combinations above

Given a sentence s and a target verb v, we want to automatically generate a set of questions containing v that are answerable with phrases from s. This task is important because generating answerable questions requires understanding the predicate-argument structure of the sentence. In essence, questions play the part of semantic roles in our approach. We present a baseline that breaks down question generation into two steps: (1) we first use a classifier to predict a set of roles for verb v that are likely present in the sentence, from a small, heuristically defined set of possibilities and then (2) generate one question for each predicted role, using templates extracted from the training set.



	Wikipedia	Wikinews	Science
Sentences	15,000	14,682	46,715
Verbs	32,758	34,026	66,653
Questions	75,867	80,081	143,388
Valid Qs	67,146	70,555	127,455

Figure 10. Question written across multiple domains

4.3.4 Alignment Pairs

A total of up to 10 *knowledge texts* are extracted with respect to each WSC problem. Each of them is processed individually along with the WSC problem to produce a corresponding intermediate result from the knowledge hunting module. This result is later used to infer the final answer to the problem.

Let $W = \langle S, A_1, A_2, P, K \rangle$ be a modified WSC problem such that S be a set of WSC sentences, A_1 and A_2 be the answer choices one and two respectively, P be the pronoun to be resolved, and K be a *knowledge text.* The existing solvers (Sharma et al. 2015) that use explicit knowledge to solve a WSC problem of the form W first convert K and S into a logical form and then use a set of axioms to compute the answer. However, it is a daunting task to convert free form text into a logical representation. Thus these methods often produce low recall. In this work, we take a detour from this approach and aim to build an "alignment" function. Informally, the task of the alignment function is to align the answer choices (A_1 and A_2) and the pronoun to be resolved P in S with the corresponding entities (noun phrases) in K. These alignments are the intermediate results of the knowledge hunting module.

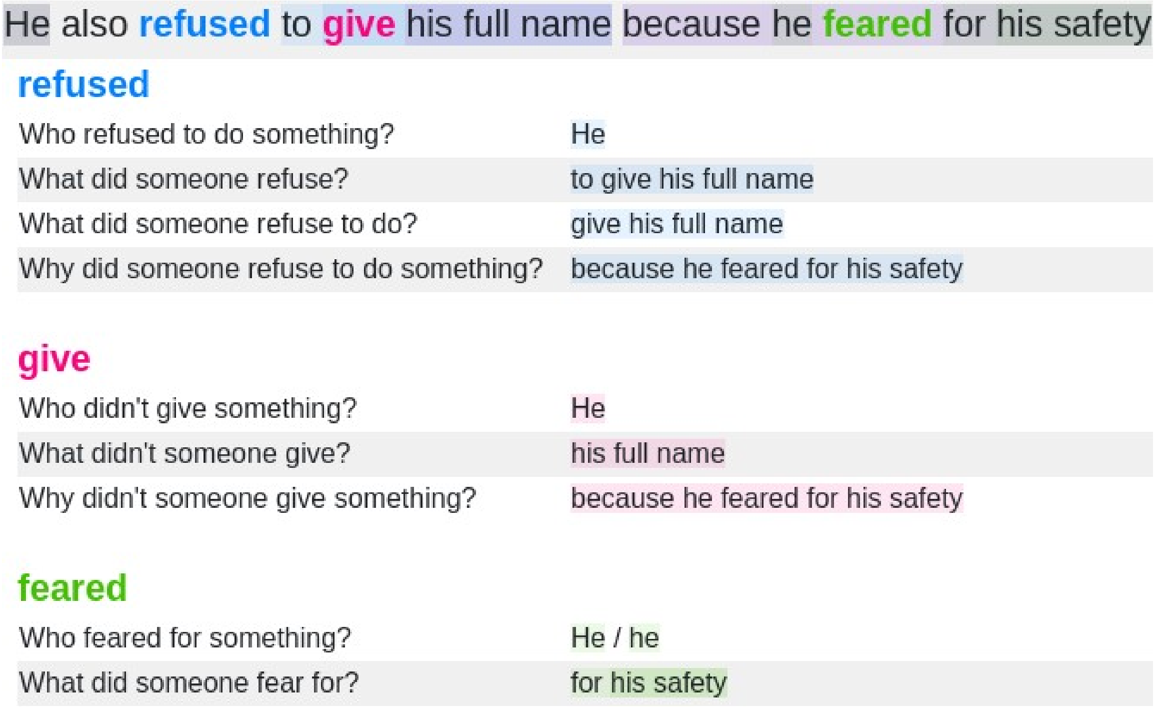
alignments are then used in a Probabilistic Soft Logic framework to infer the final answer.

Informally, the job of the alignment function is to decide which replacement $T[P/A_1]$ or $T[P/A_2]$ closely mimics K. Here, $T[P/A_i]$ represents the text that is obtained by replacing the pronoun P by A_i in T.

By the choice of knowledge extraction approach, the knowledge texts are similar to the WSC sentences in terms of events, i.e., they contain similar verb phrases, properties and discourse connectives. So, in an ideal situation we will have entities in K corresponding to each one of the concerned entities (A_1, A_2 and P) in W respectively. The goal of the alignment algorithm is to find that mapping. Let E_{k1}, E_{k2} and E_{k3} be the entities in K which correspond to the entities A_1, A_1 and P in W respectively. The mapping result is generated in the form of a *aligned_with* predicate of arity three. The first argument represents an entity (an answer choice or the pronoun) from S, the second argument represents an entity from K and the third argument is an identifier of the knowledge text used. For example, if an entity *"city councilmen"* in a WSC text aligns with an entity *"He"* in a *knowledge text* then the intermediate result We define an entity (noun phrase) E_j from a *knowledge text* K to be *aligned_with* to an entity A_j from a WSC text S if the following holds:
1) There exists a verb v in S and v' in K such that either $v = v'$ or v is a synonym of v'.
2) The "semantic role" of A_j with respect to v is same as the "semantic role" of E_j with respect to v'.

We use the semantic role labelling function, called QASRL (He, Lewis, and Zettlemoyer 2015) to compute the semantic roles of each entity. QASRL represents the semantic roles of an entity, in terms of question-answer pairs. Figure 12 shows the

QASRL representation of the *knowledge text* "*He* also refused to give his full name because *he* feared for his safety." It involves three verbs "refused", "feared" and "give". The questions represent the roles of the participating entities.



He also refused to give his full name because he feared for his safety

refused

Who refused to do something?	He
What did someone refuse?	to give his full name
What did someone refuse to do?	give his full name
Why did someone refuse to do something?	because he feared for his safety

give

Who didn't give something?	He
What didn't someone give?	his full name
Why didn't someone give something?	because he feared for his safety

feared

Who feared for something?	He / he
What did someone fear for?	for his safety

Figure 11. QASRL output for the sentence *"He also refused to give his full name because he feared for his safety."*

An example alignment generated for the WSC sentence S = *"The city councilmen refused the demonstrators a permit because they feared violence."* and the *knowledge text* K = *"He also refused to give his full name because he feared for his safety."* is *aligned_with(city councilmen,He,K), aligned_with(they,he,K).*

There are three relevant entities in an input WSC problem, i.e., A_1, A_2 and P. Based on the existence of the entities corresponding to the entities in the WSC problem there are 2^8 possible cases. For example, the case *{True True True}*, abbreviated as

{TTT}, represents that each of the entities A_1, A_2 and P are aligned with corresponding entities in a *knowledge text*.

Case	Details	Example
TTT	Each entity (among A_1, A_2 and P) in the WSC sentences W have corresponding entities in the corresponding *knowledge text* K	**WSC Sentence:** ***Jim*** *comforted* ***Kevin*** *because* ***he*** *was so upset .* **Knowledge Text (K):** *She says* ***I*** *comforted* ***her****, because* ***she*** *was so upset* **Alignments:** *aligns_with(Jim,I,K), aligns_with(Kevin,her,K), aligns_with(he,she,K)*
TFT	Only the entity representing the answer choice one (A_1) and the pronoun to be resolved (P) have corresponding entities in the *knowledge text* K	**WSC Sentence:** *The* ***trophy*** *does not fit into the brown* ***suitcase*** *because* **it** *is too large .* **Knowledge Text (K):** *installed CPU and* **fan** *would not fit in because the* ***fan*** *was too large* **Alignments:** *aligns_with(trophy,fan,K), aligns_with(it,fan,K)*
FTT	Only the entity representing the answer choice 2 (A_2) and the pronoun to be resolved (P) have corresponding entities in the *knowledge text* K	**WSC Sentence:** ***James*** *asked* ***Robert*** *for a favor but* **he** *refused .* **Knowledge Text (K):** *He asked the* ***LORD*** *what he should do, but the* ***LORD*** *refused to answer him, either by dreams or by sacred lots or by the prophets.* **Alignments:** *aligns_with(Robert,LORD,K) and aligns_with(he,LORD,K)*

Table 2. Alignment Cases in the Knowledge Hunting Approach. A_1 and A_2 are answer choices one and two, P is pronoun to resolve, E_{k1}, E_{k2} and E_{k3} are entities in a *knowledge text* (K)

The intuition behind the alignment approach is to find a common entity in a *knowledge text* such that it aligns with one of the answer choices (say A_i) and also with the pronoun to be resolved (P). Then we can say that both A_i and P refer to same entity and hence they refer to each other. An important aspect of such a scenario is the existence of the entities in a *knowledge text* which align with at least one of the answer choices and the pronoun to be resolved. In other words the cases *{TTT}*, *{TFT}* and *{FTT}*. So we consider the alignments generated only with respect to these three cases as an output of the alignment module. The three cases and their details are shown in the Table 2 along with examples from the dataset.

4.4 Language Models

In recent years, deep neural networks have achieved great success in the field of natural language processing (X. Liu et al. 2019; Chen, Li, and Xu 2018). With the recent advancements in the neural network architectures and availability of powerful machine it is possible to train unsupervised language models and use them in various tasks (Devlin et al. 2018; Trinh and Le 2018). Such language models are able to capture the knowledge which is helpful in solving many such problem sentences we have discussed so far.

S3: I put the heavy book on the table and it broke.

Pronoun to resolve: it

Answer Choices: a) table b) book

A knowledge that, *"table broke"* is more likely than *"book broke"* is sufficient to solve the above WSC problem. Such a knowledge is easily learned by the language models because they are trained on huge amounts of text snippets which are transcribed by people. The models are good at learning the frequently occurring patterns from data.

In this work, we aim to utilize such knowledge embedded in the neural language models. We replace the pronoun to be resolved in the WSC text with the two answer choices, one at a time, generating two possible texts. For example the two texts generated in the above WSC example are, S3(a) = *I put the heavy book on the table and table broke.*, S3(b) = *I put the heavy book on the table and book broke.* Then a pre-trained language model is used to predict the probability of each of the generated texts. Let P_a be the probability of S3(a) and P_b be the probability of S3(b). To be able to use the result of language models in Probabilistic Soft Logic, the output of this step contains ***coref(****P****,****A_1****):****$PROB1$* and ***coref(****P****,****A_2****):****$PROB2$*, where P is the pronoun to be resolved, A_1 and A_2 are answer choices one and two respectively, and $PROB1$ and $PROB2$ are the probabilities of the texts generated by replacing P with A_1 and A_2 in the WSC text respectively, i.e., P_a and P_b in the example above.

4.4.1 Trinh and Le's Model

In this work the author first present a simple approach for common sense reasoning with Winograd schema multiple choice questions. (Trinh and Le 2018)Key to their method is the use of language models (LMs), trained on a large amount of unlabeled data, to score multiple choice questions posed by the challenge and similar datasets. More concretely, in the above example, they first substitute the pronoun ("it") with the candidates ("the trophy" and "the suitcase"), and then use LMs to compute the probability of the two resulting sentences ("The trophy doesn't fit in the suitcase because the trophy is too big." and "The trophy doesn't fit in the suitcase because the suitcase is too big."). The below example from (Trinh and Le 2018) shows the method.

The replacing the pronoun with answer choices results in a more probable sentence will be the correct answer. The authors first replace the pronoun in the original sentence with each of the candidate choices. The problem of coreference resolution then reduces to identifying which replacement results in a more probable sentence. By reframing the problem this way the authors claim that, language modeling becomes a natural solution.

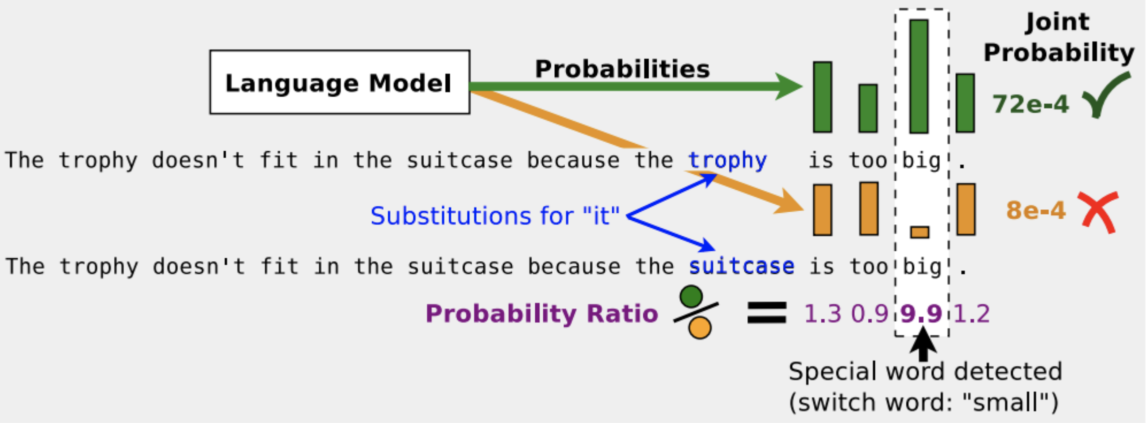


Figure 12. Substitution method with the candidate choices for the pronoun

4.4.2 Problem Representation In BERT

In BERT, we are using the next sentence prediction model. We choose do the following to break the sentences into two sentences for prediction purposes. We find the probability of the second sentence occurring given the first sentence.

1. Append *?* from the right side of sentence from *pronoun* to separate the sentence into two.
2. If a period is found split the split the where the *pronoun* exists to append *?* for splitting them.

   For example: *The painting in Mark's living room shows an oak tree. It is to the right of the bookcase.*

   In the above sentence, *"It"* is the pronoun we need resolve for choices *"The painting"* or *"The oak tree"*. It becomes the following after processing the sentences to break it for next sentence prediction:

   a) *The painting in Mark's living room shows an oak tree. ? The painting is to the right of the bookcase.*

b) *The painting in Mark's living room shows an oak tree. ? The oak tree is to the right of the bookcase.*

3. Find discourse connectives in the sentence and break it to append *. ?*

   Similarly in cases where there is no period found with definite sentences, we find the *discourse connectives* in the sentence to break it.

   For example: *Anna did a lot worse than her good friend Lucy on the test, because she had studied so hard.*

   In the above sentence, *"she"* is the pronoun we need resolve for choices *"Anna"* or *"Lucy"*. It becomes the following after processing the sentences to break it for next sentence prediction:

   a) *Anna did a lot worse than her good friend Lucy on the test. ? because anna had studied so hard.*

   b) *Anna did a lot worse than her good friend Lucy on the test. ? because lucy had studied so hard*

4.4.3 Problem Representation In RoBERTa

In RoBERTa we use the disambiguation task to find the confidence for the answer choices, we encode specifically mention the answer choices for it confidence scores like mentioned in the work of (Zhu et al. 2015).

For example: *Anna did a lot worse than her good friend Lucy on the test, because she had studied so hard.*

In the above sentence, *"she"* is the pronoun we need resolve for choices *"Anna"* or *"Lucy"*. It becomes the following after processing the sentences for the disambiguation task in RoBERTa:

1. *_Anna_ did a lot worse than her good friend Lucy on the test, because [she] had studied so hard.*
2. *Anna did a lot worse than her good friend _Lucy_ on the test, because [she] had studied so hard*

4.5 Augmenting Knowledge Module And Language Models

In this step, the alignment results generated from the knowledge hunting module and the co-reference probabilities generated from the language models are combined in a Probabilistic Soft Logic (PSL) (Kimmig et al. 2012) framework to infer the confidence for each of the choices in the problem.

4.5.1 Inference And Learning In PSL

PSL is primarily designed to support Most Probable Explanation (MPE) inference. MPE inference is the task of finding the overall interpretation (combination of grounded atoms) with the maximum probability given a set of evidence. Intuitively, the interpretation with the highest probability is the interpretation with the lowest distance to satisfaction. In other words, it is the interpretation that tries to satisfy all rules as much as possible.

We used the PSL framework to combine the results from the other modules in our approach and generate the confidence scores for each of the answer choices. The confidence scores are generated for the predicate *coref(p,a_i)* where p is the variable representing a pronoun to be resolved in a WSC problem and a_i is a variable representing an answer choice in the WSC problem.

To be able to use the alignment information from the knowledge hunting approach, following PSL rule was written. It is used to generate the *coref* predicate and its truth value for the answer choices.

$$
\begin{aligned}
w : \{\forall a, e1, e2, k, p. \\
aligned\_with(a, e1, k) \wedge \\
aligned\_with(p, e2, k) \wedge \\
similar(e1, e2) \wedge \\
& \rightarrow coref(p, a)\}
\end{aligned}
\tag{4.1}
$$

Here w is the weight of the rule, a, p, $e1$, $e2$ and k are variables such that a is an answer choice in a WSC problem, p is the pronoun to be resolved in a WSC problem, and $e1$ and $e2$ are entities in a knowledge text k. The groundings of the *aligned_with* predicate are generated from the knowledge hunting module and the groundings of the *similar* predicate encode the similar entities in k. The truth value of a grounding of *similar* predicate is used to represent how similar the two entities, i.e., $e1$ and $e2$, are to each other. Although any kind of semantic similarity calculation algorithm may be used for producing the similar predicate, we used BERT (Devlin et al. 2018) to calculate the similarity between two entities. In case the values of $e1$ and $e2$ are same (say E) the truth value of the grounded atom $similar(E, E)$ becomes 1.

Intuitively, the above rule means that if an answer choice and the pronoun to be resolved in a WSC problem align with similar entities in a knowledge text corresponding to the WSC problem then the pronoun refers to the answer choice.

The above rule is applicable to all the three cases mentioned in the Table 2.

The neural language models approach produces two groundings of the atom defined by the binary predicate *coref* as its result (see section 4.4). The two groundings refer to the co-reference between the pronoun to be resolved and the two answer choices respectively. The groundings are accompanied with their probabilities which we used as their truth values. These grounded $coref$ atoms are directly entered as input to the PSL framework along with the output from knowledge hunting approach to infer the truth values for the $coref$ atom with respect to each of the answer choices. Finally, the answer choice with higher truth value is considered as the correct co-referent of the pronoun to be resolved and hence the final answer.

Let us consider the TTT case. In case TTT, each one of the two answer choices, and the pronoun to be resolved has an entity aligned to them in a knowledge text (say $K1$). A logical rule generated for the TTT case is as shown below.

$$\begin{aligned} 0.3 : aligned\_with(A, E1, K) \wedge \\ aligned\_with(P, E2, K) \\ \wedge A \neq P \wedge similar(E1, E2) \\ \rightarrow coref(P, A) \end{aligned} \tag{4.2}$$

Here, W is the weight of the rule which is set to 0.3, the groundings of the 'aligned_with' predicate are generated from the knowledge hunting module, A represents an answer choice, P represents the pronoun to be resolved, $E1$ and $E1$ are two entities in the knowledge text K and the 'similar' predicate is used to represent how similar the two entities are to each other. Any type of similarity calculating algorithm may be used for producing the similar predicate. We used BERT (Devlin et al. 2018) to calculate the similarity.

$$\begin{aligned} 0.3 : language\_model(P, A) \\ \rightarrow coref(P, A) \end{aligned} \tag{4.3}$$

Language model scores are also given as an input to the PSL as a rule. Here P is the pronoun and A is the answer choice, for which we have confidence from the language model. This rule carries the same weight has the alignment rule. This eliminates bias for one over the other.

Intuitively, the above rule means that if an answer choice A_i in WSC is aligned with an entity E_{k1} in a knowledge text K, the pronoun to be resolved P is aligned with an entity E_{k2} in K and E_{k1} and E_{k2} are similar to each other then P refers to the answer choice A_i.

$$\begin{aligned} 0.3 : aligned\_with(A, E1, K) \wedge \\ aligned\_with(P, E2, K) \\ \wedge A \neq P \wedge not\_equal(A, P) \\ \rightarrow coref(P, A) \end{aligned} \tag{4.4}$$

Similarly to include the language model scores we use the following rule which finally gives the *cause_ of* score for a particular alternative.

$$W : language\_model(p, a) \rightarrow cause\_of(p, a) \tag{4.5}$$

For the both rules we have set the weight w to be 0.3, because both considered equal weight for solving this problem.

The logical rules for CASE FTT or TFT would look similar to the above *coref* rule in PSL. If an answer choice from either option aligns with entity E_{k1} in knowledge text K, and the choices being compared is not the same, then it forms a co reference with the pronoun given.

Chapter 5

EXPERIMENTS AND RESULTS

5.1 Experiments

5.1.1 Dataset

First, The Winograd Schema Challenge corpus[1] consists of pronoun resolution problems where a set of sentences is given along with a pronoun in the sentences and two possible answer choices such that only one choice is correct. There are 285 problems in the WSC dataset. From this point onward, we will call this dataset as WSC_{285}. The generation of the original WSC dataset itself is an ongoing work. Hence the dataset keeps getting updated. This is why the works earlier than ours, used a smaller dataset containing 273 problems. All the problems in it are also present in WSC_{285}. From this point onward, we will call this subset of WSC_{285} as WSC_{273}. For a fair comparison between our work and others', we performed our experiments with respect to both WSC_{285} and WSC_{273}.

Second, The Cause of Plausible Alternatives consists of alternatives, where is a premises and two possible hypothesis is given and only one hypothesis is correct. There are totally 1000 problem statements, in which 500 train and 500 test sets are provided. From this point let's called the train set has $COPA_{TRAIN}$ and test set has $COPA_{TEST}$. The experiments were performed on with the BERT model to compare

[1] Available at https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WSCollection.xml

the performance of METHOD$_{BERT}$. The results on this dataset are shown in the below table.

To test the same on a much larger scale, experiments where also conducted on Winogrande. Which is larger dataset of the Winograd. This contains 38000 training set, 4279 test set and 2863 validation set. We will call the sets as $GRANDE_{TRAIN}$, $GRANDE_{TEST}$ and $GRANDE_{VAL}$ respectively.

5.1.2 Experimental Setup And Results

To start with, we compared the results of the previous systems to our system. We compared the number of predictions made in a particular problem. These predictions are compared with the use their precision, recall and F1 Scores. The language models based component of our approach relies on pre-trained language models. Here we compared three different language models. First we used the ensemble of 14 pre-trained language models which are used in (Trinh and Le 2018). Secondly, we used BERT (Devlin et al. 2018) pre-trained model and for the last we used RoBERTa (Y. Liu et al. 2019). Based on the language model used, in the following experiments we use METHOD$_{T2018}$ to represent our approach which uses models from (Trinh and Le 2018), METHOD$_{BERT}$, and METHOD$_{RoBERTa}$ to represent our approach which uses the BERT language model. We compared our method with three language models with datsets of WSC, COPA and Winogrande.

5.1.2.1 Winograd Experiments Setup

The comparison results are as shown in the Table 3. The first two, (Sharma et al. 2015) and (Q. Liu et al. 2017) hereafter called S2015 and L2017 respectively, address a subset of WSC problems (71 problems). Both of them are able to exploit only causal knowledge. This explains their low coverage over the entire corpus. We overcome this issue by using any form of knowledge text making predictions for each of the problems in the dataset. More recently, two approaches on solving the WSC_{273} dataset have been proposed. The first work (Emami, De La Cruz, et al. 2018) (hereafter called E2018) extract knowledge in form of sentences to find evidences to support each of the possible answer choices. We have demonstrated both the systems and presented scores, moreover we have combined the system to check performance with our approach and compared their results. A comparison between their results and our approach can be seen the two tables 3 and 5.

We use the knowledge hunting module, which is an information retrieval process using the search engine with search queries. Search queries are formed using the part of speech tagging methods available in Stanford NTLK (Toutanova and Manning 2000). Web search is done on Google, and the sentences are parsed using a web crawler.

After retrieving the information we use the knowledge ranking method to score them. Here we have experimented with the ELMO entailment (Peters et al. 2017) and the BERT STS methods (Reimers and Gurevych 2019). We compare both the scores in table 5. We use the language models used in the systems of (Trinh and Le 2018), (Devlin et al. 2018) and (Y. Liu et al. 2019) to compare their results after augmenting their scores with knowledge using PSL.

	#correct	**% Correct**	**Correct Pairs**	**Incorrect Pairs**	**#Times Choice2 is Chosen**
S2015	49	18.0	-	-	-
L2017	43	15.0	-	-	-
E2018	119	44.0	-	-	-
T2018 (WSC_{273})	174	63.70	42	89	142
T2018 (WSC_{285})	180	63.15	44	97	146
BERT Only (WSC_{273})	173	63.36	36	94	129
BERT Only (WSC_{285})	179	58.60	-	-	-
BERT Only ($COPA_{TEST}$)	293		37	101	131
ROBERTA Only (WSC_{285})	258	90.5	116	25	-
BERT Only ($GRANDE_{VAL}$)	1631	56.98	473	958	-
BERT$_{FT}Only$(GRANDE$_{VAL}$)	1882	65.74	565	866	-
ROBERTA Only ($GRANDE_{VAL}$)	2163	75.61	702	729	-

Table 3. Results on the previous approach

Works (Trinh and Le 2018) (hereafter called T2018) uses a neural network architecture to learn language models from huge data sources to predict the probability of choosing one answer over the other. By breaking the sentences into two, we use the BERT based Next Sentence Prediction model for identifying the probability of the choices. Finally, RoBERTa uses a disambiguation model which masks the words in the problem sentence to identify the probability choices. These method are also compared and their results are shown in the above table 5

	#correct	**% Correct**	**Correct Pairs**	**Incorrect Pairs**	**#Times Choice2 is Chosen**
METHOD$_{T2018}$ (WSC_{273})	**189**	**69.23**	60	71	143
METHOD$_{T2018}$ (WSC_{285})	**195**	**68.42**	61	80	148
METHOD$_{BERT}$ (WSC_{273})	**194**	**71.06**	57	74	130
METHOD$_{BERT}$ (WSC_{285})	**200**	**70.17**	58	83	134
METHOD$_{ROBERTA}$ (WSC_{285})	**259**	**90.8**	115	26	-
METHOD$_{BERT}$ ($COPA_{TEST}$)	**306**	**61.2**	-	-	-
METHOD$_{BERT}$ ($GRANDE_{VAL}$)	**1729**	**60.39**	535	896	-
METHOD$_{BERT_{FT}}$ ($GRANDE_{VAL}$)	**2036**	**71.11**	651	780	-
METHOD$_{ROBERTA}$ ($GRANDE_{VAL}$)	**2211**	**77.2**	743	688	-

Table 4. Results on the proposed method

	#correct	**% Correct**	**Correct Pairs**	**Incorrect Pairs**
RoBERTa Only (WSC_{273})	247	90.47	116	20
RoBERTa Only (WSC_{285})	248	87.01	116	25
RoBERTa Only ($GRANDE_{VAL}$)	2163	75.55	702	729
METHOD$_{R}oBERTa$(WSC$_{273}$)	252	92.3	117	19
METHOD$_{R}oBERTa$(WSC$_{285}$)	253	88.77	117	24
METHOD$_{R}oBERTa$(GRANDE$_{VAL}$)	2194	76.63	674	757

Table 5. Our Approach with RoBERTa using BERT STS for similarity

We performed a second set of experiments to further investigate the robustness of our method as compared to the state of the art system (T2018). Each problem in the WSC has a sister problem in the WSC such that the texts in the two problems differ only by a word or two but the same pronoun refers to different entities. The two answer choices for both the problems in the pair are also same. For example, consider the following pair of problems.

S4: The firemen arrived **after** the police because they were coming from so far away.
Pronoun to resolve: they
Answer Choices: a) firemen b) police

S5: The firemen arrived **before** the police because they were coming from so far away .
Pronoun to resolve: they
Answer Choices: a) firemen b) police

In the above problems, only changing one word (*before/after*) in the sentence changes the answer to the problem. Due to this property of the dataset, a system can achieve an accuracy of 50% by just answering choice 1 as the correct answer for every problem. To make sure that this is not the case in our system, we performed the following two experiments.

- **Experiment to Evaluate Pairwise Accuracy:** In this experiment we evaluate our method and the other methods to find out how many of the problem pairs were correctly solved. The table 3 shows the results of the experiment. It can be seen from the results that our best performing method(METHOD$_{BERT}$ on WSC$_{273}$) solves 57 pairs correctly, which is significantly more than its baseline 'BERT Only' method. Similar pattern for the other methods can be seen in the Table 3.

- **Experiment to Evaluate System Bias:** In this experiment we evaluate our method and the others to find out if the methods are biased to chose the answer choice which is closer to the pronoun in a WSC sentence. We found that usually the answer choice 2 in the problem is closer to the pronoun to be resolved. Hence the experiments were performed to figure out how many times a method answers choice 2 as the final answer. The results of the experiments are as shown in the Table 3. As seen from the results, both, the language model based methods and our methods are not particularly biased towards one of the answer choices.

5.1.2.2 Winogrande Experiments Setup

We follow the same methods as we mentioned in section 5.1.2.1. Since the dataset size is 43000, we had the capacity to fine tune the language model using the pre-trained information. The fine tuning on the system (Devlin et al. 2018) is done, their results are compared against pretrained and fine tuned. Using the RoBERTa disambiguation model along with our appraoch we were able to compare the result between the disambiguation model and our method. The experiments were conducted using the validation set of the problem, since the test set answers were not available when the experiments are performed.

5.1.2.3 COPA Experiments Setup

Experiment on COPA is was chosen for the purpose of displaying that the method could work on such similar problems. Problems which required commonsense knowledge can use this method to improve. We used the similar method as employed in the section 5.1.2.1, but we did this knowledge retrieval for both the alternatives in order to collect evidences for both *CAUSE* and *EFFECT* examples. We use the system of (Devlin et al. 2018) and augment them with the knowledge using PSL to get the final confidence.

5.1.3 Remarks

An important contribution of our work is the improvement on the results of a language model. Our approach does that by automatically extracting the suitable knowledge.

5.1.3.1 Evaluation

When we compared (METHOD$_{BERT}$ on WSC$_{273}$), it correctly answers 26 problems which are incorrectly answered by the baseline language model (BERT Only on WSC$_{273}$). We found that the main reason for such a behavior is the addition of the suitable knowledge from the knowledge hunting module. It helps in generating the support for the correct answer to the extent that it overturns the decision of the language model. For example, we observed that for the WSC sentence

S4: The woman held the girl against her will.

Pronoun to resolve: her

Answer Choices: a)the woman b) the girl

'The woman held the girl against her will' the BERT language model predicted that *'her'* refers to *'The woman'* with the probability score of 0.513, which is incorrect, and to *'the girl'* with the probability score of 0.486. But the knowledge hunting approach alone within the PSL framework predicted the answer to be *'the girl'* with the probability score of 0.966, which is correct, and the answer *'the woman'* with the probability score of 0.034. Overall the PSL inference engine combined scores from both the approaches and corrected the decision made by the language model by predicting *'the girl'* as the correct answer with the probability score of 0.967.

On the other hand five problems were found to be incorrectly answered by our approach which were correctly answered by the language model. In all such cases the probabilities corresponding to the answer choices were found to be very close to each other and inclining towards the incorrect answer. The difference between language model probabilities generally being very small, the combined approach answered incorrectly in such cases. The main reason for such a behavior is the availability of unsuitable *knowledge text*. For example the *knowledge text* for the WSC sentence *'The man lifted the boy onto his shoulders .'* was *'If she scores I'll feel really bad!' New documentary lifts the lid on life for female stars who are partners but line up for rival clubs'*.

In similar method, when (Y. Liu et al. 2019) is replaced for the language model scores, on the WSC_{273}, it was able to solve *90.5%* of the problems, after using our approach we were able to achieve **92.3%**. It was able to solve 5 extra problems, for the similar reasons mentioned in section 5.1.3.1 in our best performance setting. In the best performance setting, we use the BERT STS model for the knowledge ranking and text similarity.

For example, the below problem

S4: Anne gave birth to a daughter last month . She is a very charming baby .

Pronoun to resolve: She

Answer Choices:

a) Anne

b) the daughter

We have compared against both the ELMO and the BERT STS methods. Below are the list of knowledge texts retrieved through the search.

1. *Earth gave birth to her last and most frightful offspring, a creature more terrible than any that had gone before.*
2. *A 600-pound woman has given birth to a 40-pound baby at Perth's King Edward Memorial Hospital.*
3. *Broadcaster Steph McGovern has given birth to a baby girl, admitting today that she is now*
4. *Freddy decided to carry his own baby after wanting to start a family, but he faced a highly unusual challenge*
5. *A day after finding out she's pregnant, a Las Vegas woman gives birth to a baby boy*
6. *These females do not lay eggs; they give birth to young aphids, all of which are females.*

7. *Steph McGovern has given birth to a baby girl, she revealed today*
8. *Police are investigating the death of a baby in Britain's largest female prison after an inmate gave birth alone in her cell at night*
9. *The tale of a woman named Catherine Bridge, who supposedly "holds the World Record for the most babies in a lone pregnancy by giving birth*
10. *A 74-Year-Old Woman Has Given Birth to Twins. Here's How That's Possible*
11. *Woman gives birth to son and then gives birth again one month later to twins*
12. *when a baby is born, it comes out of its mother's body and starts its life.*
13. *The woman, surnamed Tian, gave birth to a girl on Friday at Zaozhuang Maternity and Child Health Hospital*
14. *Well hello world, just surfaced to let you know that we now have a daughter*
15. *Verizon surprised by lawsuit, says it already fixed the problem.*
16. *Activists objected to plans to give Amazon almost $3 billion in incentives.*

Ranking through the ELMO entailment method, we pick the top 10 knowledge text for use.

1. *These females do not lay eggs; they give birth to young aphids, all of which are females.*
2. *A 74-Year-Old Woman Has Given Birth to Twins. Here's How That's Possible*
3. *Earth gave birth to her last and most frightful offspring, a creature more terrible than any that had gone before.*
4. *Broadcaster Steph McGovern has given birth to a baby girl, admitting today that she is now*
5. *A day after finding out she's pregnant, a Las Vegas woman gives birth to a baby boy*
6. *Steph McGovern has given birth to a baby girl, she revealed today*

7. *Woman gives birth to son and then gives birth again one month later to twins*
8. *The tale of a woman named Catherine Bridge, who supposedly "holds the World Record for the most babies in a lone pregnancy by giving birth*
9. *Police are investigating the death of a baby in Britain's largest female prison after an inmate gave birth alone in her cell at night*
10. *Activists objected to plans to give Amazon almost $3 billion in incentives.*

In BERT STS method, the same method is followed to rank. This uses the similarity scores

1. *Freddy decided to carry his own baby after wanting to start a family, but he faced a highly unusual challenge*
2. *These females do not lay eggs; they give birth to young aphids, all of which are females*
3. *Earth gave birth to her last and most frightful offspring, a creature more terrible than any that had gone before.*
4. *A 600-pound woman has given birth to a 40-pound baby at Perth's King Edward Memorial Hospital.*
5. *Woman gives birth to son and then gives birth again one month later to twins*
6. *A 74-Year-Old Woman Has Given Birth to Twins. Here's How That's Possible*
7. *A day after finding out she's pregnant, a Las Vegas woman gives birth to a baby boy*
8. *Police are investigating the death of a baby in Britain's largest female prison after an inmate gave birth alone in her cell at night*
9. *Well hello world, just surfaced to let you know that we now have a daughter*
10. *when a baby is born, it comes out of its mother's body and starts its life.*

In the ELMO method, the ranking is different when compared with the BERT STS method. In the ELMO method, the knowledge text has also included *Activists objected to plans to give Amazon almost $3 billion in incentives.*, which is not relevant knowledge to the problem sentence. When QASRL is ran for the same text, we get the following.

Sentence: Anne gave birth to a daughter last month . She is a very charming baby .

Who V Something? Anne

Knowledge Text: Activists objected to plans to give Amazon almost $3 billion in incentives.

Who V Something? Activists

There is aligned formed between *Activists* and *Anne*, the evidence for choice *Anne* is strengthened because of this reason. When BERT STS is used, this particular knowledge is excluded which solves the problem in this setting. More relevant knowledge texts are ranked through this method, which increases the accuracy in WSC$_{273}$.

When the same is experimented on the $GRANDE_{VAL}$, we get quite different results. The knowledge texts are reordered and some are excluded in the case of $GRANDE_{VAL}$. To improve this, the span of the knowledge texts were increased to 20. But doing this again reduced the accuracy of the approach, since irrelevant evidences were added to the wrong choices. Knowledge extraction and ranking method still needs to be improved for better results.

5.1.3.2 Error Analysis

Analysis this method is done on the best performing mode, since *METHOD_RoBERTa* is the best performing method, the examples below presented are for the same. Evaluation is done on the WSC dataset to get in much more details. This method fails to only solve 21 problems in the WSC dataset. With careful analysis on the same, we were able to see that needed knowledge to solve these problems are inadequate in the system.

We present some of the problems which RoBERTA and PSL predicted wrongly in the following table 6

Problem	**Answer**	**Knowledge Present**	**LM Score**	**PSL Output**
The lawyer asked the witness a question. but [he] was reluctant to answer it	**Choice1** The Lawyer **Choice2** The Witness **Answer** *the witness*	No	**Choice1** 0.521 **Choice2** 0.513	**Choice1** 0.503 **Choice2** 0.495
Sam's drawing was hung just above Tina's. and [it] did look much better with another one above it .	**Choice1** Sam's Drawing **Choice2** Tina's **Answer** *Tina's*	Yes	**Choice1** 0.528 **Choice2** 0.443	**Choice1** 0.513 **Choice2** 0.475
Alice tried frantically to stop her daughter from barking at the party leaving us to wonder. why [she] was behaving so strangely .	**Choice1** Alice **Choice2** Daughter **Answer** *Daughter*	Yes	**Choice1** 0.523 **Choice2** 0.447	**Choice1** 0.538 **Choice2** 0.490
I could not find a spoon, so I tried using a pen to stir my coffee. But that turned out to be a bad idea ,. because [it] got full of ink .	**Choice1** The pen **Choice2** The Coffee **Answer** *The coffee*	No	**Choice1** 0.370 **Choice2** 0.302	**Choice1** 0.550 **Choice2** 0.448
The sun was covered by a thick cloud all morning . but luckily , by the time the picnic started , [it] was gone .	**Choice1** The Sun **Choice2** Cloud **Answer** *Cloud*	No	**Choice1** 0.613 **Choice2** 0.216	**Choice1** 0.529 **Choice2** 0.469

Table 6. Incorrect prediction by our approach and RoBERTa

As we can see in the table, the difference for the actual answer score have increase with knowledge text, while the scores didn't change much when there is no knowledge. For example in the examples given above *Sam's drawing was hung just above Tina's. and [it] did look much better with another one above it* , we can see that the answer is

Tina's drawing. From RoBERTa we get the score of **0.528** and **0.443** for choice1 and choice2 respectively. After augmenting the knowledge and calculating the final score with PSL, we get **0.513** and **0.475**. Even though we got the wrong prediction with the higher score for choice1, choice2 score from the PSL prediction has slightly increased when compared with the RoBERTa's score. This shows when we add knowledge to solve the problem, more evidences can predict the answer correctly.

We also did analysis on RoBERTa predict correctly and PSL wrongly predicts them.

Problem	**Answer**	**Reason**	**LM Score**	**PSL Output**
Sam and Amy are passionately in love , but Amy's parents are unhappy about it . because [they] are snobs .	**Choice1** Sam and Amy **Choice2** Amy's Parents **Answer** *Amy's Parents*	Irrelevant Knowledge	**Choice1** 0.357 **Choice2** 0.418	**Choice1** 0.997 **Choice2** 0.539
Fred is the only man alive who still remembers my father as an infant . when Fred first saw my father , [he] was twelve months old.	**Choice1** Fred **Choice2** My Father **Answer** *Fred*	Knowledge Aligned to Choice1	**Choice1** 0.543 **Choice2** 0.571	**Choice1** 0.983 **Choice2** 0.710
In July , Kamtchatka declared war on Yakutsk . Since Yakutsk's army was much better equipped and ten times larger , [they] were defeated within weeks .	**Choice1** Kamchatka **Choice2** Yakutsk **Answer** *Kamchatka*	Knowledge Aligned to Choice2	**Choice1** 0.341 **Choice2** 0.331	**Choice1** 0.960 **Choice2** 0.967
The man lifted the boy onto [his] shoulders .	**Choice1** The man **Choice2** The boy **Answer** *The man*	Knowledge Aligned to Choice2	**Choice1** 0.464 **Choice2** 0.456	**Choice1** 0.966 **Choice2** 0.983

Table 7. Correct predictions by RoBERTa, but incorrect by our approach

In the above table, examples are shown were our approach wrongly classified. We could see that, irrelevant knowledge or wrong alignments made wrong predictions. In the case of irrelevant knowledge, there could be alignments for both the choices, since there are more knowledge texts, the evidences could randomize the predictions, thus resulted in the wrong choice. Similarly with knowledge aligned to one particular choice, the knowledge text had wrong alignment with the one of the choices to predict that to be the answer. There were 4 such cases found when using the RoBERTa.

In the below table, it is shown some example where RoBERTa wrongly predicted and our approach predicted correctly.

Problem	**Answer**	**Reason**	**LM Score**	**PSL Output**
Frank was upset with Tom. because the toaster [he] had bought from him did not work	**Choice1** Frank **Choice2** Tom **Answer** *Frank*	Relevant Knowledge	**Choice1** 0.531 **Choice2** 0.460	**Choice1** 0.983 **Choice2** 0.462
The older students were bullying the younger ones . so we rescued [them] .	**Choice1** The older ones **Choice2** The younger ones **Answer** *The younger ones*	Relevant Knowledge	**Choice1** 0.440 **Choice2** 0.432	**Choice1** 0.503 **Choice2** 0.966
Bob paid for Charlie's college education . [He] is very grateful .	**Choice1** Bob **Choice2** Charlie **Answer** *Charlie*	Relevant Knowledge	**Choice1** 0.494 **Choice2** 0.370	**Choice1** 0.818 **Choice2** 0.966
Mary tucked her daughter Anne into bed ,. so that [she] could sleep .	**Choice1** Mary **Choice2** Anne **Answer** *Anne*	Relevant Knowledge	**Choice1** 0.472 **Choice2** 0.397	**Choice1** 0.544 **Choice2** 0.966

Table 8. Incorrect predictions by RoBERTa, but predicted correctly by our approach

There were totally 9 problems which were correctly predicted by our approach which was wrongly predicted by RoBERTa. In table 8, some examples are presented along with predicted scores in the systems. We could see that the strong evidence from the knowledge texts were able to outperform the RoBERTa's scores. This shows that relevant knowledge to the problem can help solve the problem with much better accuracy.

The complete set of knowledge hunting modules, language models and PSL rules are available at https://github.com/Ashprakash/CKLM

Chapter 6

DISCUSSION

6.1 Conclusion

Automatic extraction of the needed commonsense knowledge is a major obstacle in solving the ambiguity in the language was well proven by solving Winograd Schema Challenge and Cause of Plausible Alternatives. It is observed that sometimes the needed knowledge is based on correlation between concepts and it can be retrieved from the pre-trained language models. At other times a more involved knowledge about actions and properties is needed. So, in this work we utilized the knowledge embedded in the pre-trained language models and developed a technique to automatically extract the actions and properties based commonsense knowledge from text repositories. Then we defined an approach to combine the two kinds of knowledge in a probabilistic soft logic based framework to solve the Winograd Schema Challenge (WSC). The experimental results show that the combined approach possesses the benefits of both the approaches and achieves the state of the art accuracy on WSC.

The experiments on COPA which deals with Causal reasoning shows us to solve problems of similar kind, knowledge is a significant component. Causal reasoning is an important task in figuring out relationship between the cause and effect. The approach we followed above gets the required knowledge to do the above reasoning. This approach gets the most relevant knowledge for the cause or effect given in the problem statement to finally derive the confidence for alternatives provided.

Though models which are trained on large corpus like Roberta performs good these data sets, Winogrande, an adversarial data sets proves it can be still faulty if new data is provided to the language model for prediction. The key take away after multiple experiments with language models shows that they are not consistent across different problems and semantics are not preserved as opposed the co-occurring words.

6.2 Recommendations

- Works are required to improve the search query generation and information retrieval to extract the needed knowledge for the problem sentence.

- Knowledge hunting module and use of such knowledge with the problem can be done through the system presented in (Sap et al. 2018). The approach is named *ATOMIC* which follows a if-then-else knowledge representation using machine commonsense.

- Approach can be applied to Question Answering problems which requires commonsense knowledge.

- Machine translation has ambiguities which prevails in a natural language. To ensure the semantics of the information is preserved while conversion, it is significant to save the commonsense extracted for the problem. These methods can be employed to preserve such knowledge for conversion.

- Improvements are required on implementing a general framework, which can distinguish between a relevant knowledge and irrelevant knowledge. Irrelevant knowledge can cause error in alignments while relevant knowledge form good alignments, thus causing better predictions.

- Finally the combination of multiple systems require probabilistic approach. A learning technique could be employed if the parameters for combination increases.

Since the parameters in this approach is an alignment pair and a language model score, PSL could derive a confidence. However if the parameters increases learning methods like machine learning or deep learning can be used to augment the knowledge.

REFERENCES

Allen, J. 1995. *Natural language understanding (2nd ed.),* vol. ISBN:0-8053-0334-0.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. “Identifying relations for open information extraction.” In *EMNLP,* 1:1–7.

Bailey, Dan, Amelia Harrison, Yuliya Lierler, Vladimir Lifschitz, and Julian Michael. 2015. “The winograd schema challenge and reasoning about correlation.” In *In Working Notes of the Symposium on Logical Formalizations of Commonsense Reasoning.*

Bailey, Daniel, Amelia Harrison, Yuliya Lierler, Vladimir Lifschitz, and Julian Michael. 2015. “The Winograd Schema Challenge and Reasoning about Correlation.” In *Logical Formalizations of Commonsense Reasoning: Papers,* 1:1–8.

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015a. “A large annotated corpus for learning natural language inference.” In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics.

———. 2015b. “A large annotated corpus for learning natural language inference.” In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,* 632–642. Lisbon, Portugal: Association for Computational Linguistics, September. doi:10.18653/v1/D15-1075.

Carbonell, Jaime G, and Ralf D Brown. 1988. “Anaphora resolution: a multi-strategy approach.” In *Proceedings of the 12th conference on Computational linguistics-Volume 1,* 96–101. Association for Computational Linguistics.

Chen, Yongrui, Huiying Li, and Zejian Xu. 2018. “Convolutional Neural Network-Based Question Answering Over Knowledge Base with Type Constraint.” In *China Conference on Knowledge Graph and Semantic Computing,* 28–39. Springer.

Clark, Peter, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. “Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge.” *arXiv preprint arXiv:1803.05457.*

Davis, Ernest, and Gary Marcus. 2015a. “Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence,” Vol 58:9.

Davis, Ernest, and Gary Marcus. 2015b. “Commonsense reasoning and commonsense knowledge in artificial intelligence.” In *Communications of the ACM,* vol. Volume 58 Issue 9, September 2015, 92–103.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. “Bert: Pre-training of deep bidirectional transformers for language understanding.” *arXiv preprint arXiv:1810.04805.*

Emami, Ali, Noelia De La Cruz, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2018. “A Knowledge Hunting Framework for Common Sense Reasoning.” In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* 1949–1958.

Emami, Ali, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2018. “A Generalized Knowledge Hunting Framework for the Winograd Schema Challenge.” In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop,* 25–31.

Gelfond, M., and Y. Kahl. 2014. *Knowledge Representation, Reasoning, and the Design of Intelligent Agents: The Answer-Set Programming Approach.* Cambridge University Press. https://books.google.com/books?id=kBL7AgAAQBAJ.

Gokaslan, Aarpn, and Vanya Cohen. 2019. “Openwebtext corpus.” http://web.archive.org/%20save/http://Skylion007.github.io/%20OpenWebTextCorpus.

Gordon, Andrew S., Zornitsa Kozareva, and Melissa Roemmele. 2012. “SemEval-2012 Task 7: Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning.” In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012).* Montreal, Canada, June. http://ict.usc.edu//pubs/SemEval-2012%20Task%207-%20Choice%20of%20Plausible%20Alternatives-%20An%20Evaluation%20of%20Commonsense%20Causal%20Reasoning.pdf.

He, Luheng, Mike Lewis, and Luke Zettlemoyer. 2015. “Question-answer driven semantic role labeling: Using natural language to annotate natural language.” In *Proceedings of the 2015 conference on empirical methods in natural language processing,* 643–653.

Isaak, Nicos, and Loizos Michael. 2016. “Tackling the Winograd Schema Challenge Through Machine Logical Inferences.” In *STAIRS,* 284:75–86.

Johnson-Laird, P. N., and G. A. Miller. 1976. *Language and Perception,* vol. The Belknap Press of Harvard University Press, Cambridge, Massachusetts.

Kimmig, Angelika, Stephen H Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2012. “A short introduction to probabilistic soft logic.” In *NIPS Workshop on probabilistic programming: Foundations and applications,* 1:3.

Levesque, Hector J, Ernest Davis, and Leora Morgenstern. 2011. “The Winograd schema challenge.” In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning,* 46:47.

Liu, Quan, Hui Jiang, Andrew Evdokimov, Zhen-Hua Ling, Xiaodan Zhu, Si Wei, and Yu Hu. 2017. “Cause-Effect Knowledge Acquisition and Neural Association Model for Solving A Set of Winograd Schema Problems.” In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI),* 2344–2350.

Liu, Xiaodong, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. “Multi-Task Deep Neural Networks for Natural Language Understanding.” *arXiv preprint arXiv:1901.11504.*

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. “RoBERTa: A Robustly Optimized BERT Pretraining Approach.” *CoRR* abs/1907.11692. arXiv: 1907.11692. http://arxiv.org/abs/1907.11692.

Marcus, G., and E. Davis. 2019. *Rebooting AI: Building Artificial Intelligence We Can Trust.* Knopf Doubleday Publishing Group. https://books.google.com/books?id=O8muDwAAQBAJ.

Melissa Roemmele1, Cosmin Adrian Bejan2, and Andrew S. Gordon2. 2011. “Choice of Plausible Alternatives (COPA) An evaluation of commonsense causal reasoning.” In *AAAI,* vol. 1, March 21-23, 1–7.

Mihaylov, Todor, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. “Can a suit of armor conduct electricity? a new dataset for open book question answering.” *arXiv preprint arXiv:1809.02789.*

Miller, George A. 1995. “WordNet: a lexical database for English.” *Communications of the ACM* 38 (11): 39–41.

Mishra, Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. 2018. “Tracking State Changes in Procedural Text: A Challenge Dataset and Models for Process Paragraph Comprehension.” *arXiv preprint arXiv:1805.06975.*

Ng, Vincent. 2017. “Machine Learning for Entity Coreference Resolution: A Retrospective Look at Two Decades of Research.” In *AAAI,* 4877–4884.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2017. “Deep contextualized word representations,” 1:1–15.

Prakash, Ashok, Arpit Sharma, Arindam Mitra, and Chitta Baral. 2019. “Combining Knowledge Hunting and Neural Language Models to Solve the Winograd Schema Challenge.” In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* 6110–6119. Florence, Italy: Association for Computational Linguistics, July. doi:10.18653/v1/P19-1614.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. “Language Models are Unsupervised Multitask Learners,” 1:1–8.

Raghunathan, Karthik, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. “A multipass sieve for coreference resolution.” In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing,* 492–501. Association for Computational Linguistics.

Reimers, Nils, and Iryna Gurevych. 2019. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.* arXiv: 1908.10084 `[cs.CL]`.

Robert Speer, Catherine Havasi, and Henry Lieberman. 2008. “AnalogySpace: Reducing the Dimensionality of Common Sense Knowledge.” In *AAAI,* 1:1–7.

Roni Khardon, Dan Roth, and Leslie G. Valiant z. 1999. “Relational Learning for NLP using Linear Threshold Elements.” In *IJCAI,* 1:1–7.

Sakaguchi, Keisuke, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. “WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale.” *CoRR* abs/1907.10641. arXiv: 1907.10641. http://arxiv.org/abs/1907.10641.

Sap, Maarten, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2018. “ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning.” *CoRR* abs/1811.00146. arXiv: 1811.00146. http://arxiv.org/abs/1811.00146.

Schüller, Peter. 2014. “Tackling winograd schemas by formalizing relevance theory in knowledge graphs.” In *Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning.*

Sharma, Arpit, Nguyen Ha Vo, Somak Aditya, and Chitta Baral. 2015. “Towards Addressing the Winograd Schema Challenge-Building and Using a Semantic Parser and a Knowledge Hunting Module.” In *IJCAI,* 1319–1325.

Singh, Push, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. “Open Mind Common Sense: Knowledge Acquisition from the General Public.” In *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE,* edited by Robert Meersman and Zahir Tari, 1223–1237. Berlin, Heidelberg: Springer Berlin Heidelberg.

Soderland, Stephen, John Gilmer, Robert Bart, and Oren Etzioni. 2013. “Open information extraction to KBP relations.” In *In Text Analysis Conference,* 1:1–7.

Toutanova, Kristina, and Christopher D. Manning. 2000. “Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-speech Tagger.” In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13,* 63–70. EMNLP ’00. Hong Kong: Association for Computational Linguistics. doi:10.3115/1117794.1117802.

Trinh, Trieu H, and Quoc V Le. 2018. “A Simple Method for Commonsense Reasoning.” *arXiv preprint arXiv:1806.02847.*

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” *CoRR* abs/1706.03762. arXiv: 1706.03762. http://arxiv.org/abs/1706.03762.

Zhu, Yukun, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. “Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books.” *CoRR* abs/1506.06724. arXiv: 1506.06724. http://arxiv.org/abs/1506.06724.