Improving Natural Language Inference Using External Knowledge in the Science Questions Domain

Xiaoyan Wang,[§] Pavan Kapanipathi,[†] Ryan Musa,[†] Mo Yu,[†]
Kartik Talamadupula,[†] Ibrahim Abdelaziz,[†] Maria Chang,[†]
Achille Fokoue,[†] Bassem Makni,[†] Nicholas Mattei,[†] Michael Witbrock[†]
[§]Department of Computer Science, University of Illinois at Urbana-Champaign
Urbana-Champaign, IL, USA xiaoyan5@illinois.edu
[†]IBM T.J. Watson Research Center, IBM Research, Yorktown Heights, NY, USA
{ramusa, kapanipa, yum, krtalamad, achille, witbrock}@us.ibm.com
{ibrahim.abdelaziz1, maria.chang, bassem.makni, n.mattei}@ibm.com

Abstract

Natural Language Inference (NLI) is fundamental to many Natural Language Processing (NLP) applications including semantic search and question answering. The NLI problem has gained significant attention due to the release of large scale, challenging datasets. Present approaches to the problem largely focus on learning-based methods that use only textual information in order to classify whether a given premise entails, contradicts, or is neutral with respect to a given hypothesis. Surprisingly, the use of methods based on structured knowledge – a central topic in artificial intelligence – has not received much attention vis-a-vis the NLI problem. While there are many open knowledge bases that contain various types of reasoning information, their use for NLI has not been well explored. To address this, we present a combination of techniques that harness external knowledge to improve performance on the NLI problem in the science questions domain. We present the results of applying our techniques on text, graph, and text-and-graph based models; and discuss the implications of using external knowledge to solve the NLI problem. Our model achieves close to state-of-the-art performance for NLI on the SciTail science questions dataset.

1 Introduction

Natural Language Inference (NLI) – also known as textual entailment – is a fundamental task in Natural Language Understanding (NLU) (MacCartney and Manning 2009). Progress on the NLI problem has been shown to improve performance on important tasks that require NLU, including semantic search, question answering, and text summarization. In accordance with this, multiple large-scale datasets for NLI have been released (Bowman et al. 2015; Khot, Sabharwal, and Clark 2018; Williams, Nangia, and Bowman 2018), and the task has garnered significant attention from the NLP community (Parikh et al. 2016; Zhao, Huang, and Ma 2016; Khot, Sabharwal, and Clark 2018; Yin, Roth, and Schütze 2018; Chen et al. 2018).

The main goal in the NLI problem is to determine whether a given natural language *hypothesis* h can be inferred from a natural language *premise* p. NLI is often cast as a classification problem: given two sentences – hypothesis and premise – the problem lies in classifying the relationship between them into one of three classes: 'entailment', 'contradiction', or 'neutral'. In this paper, we restrict our focus in solving the NLI problem to the "entailment" class, as it is the most salient to down-stream tasks such as question answering (Khot, Sabharwal, and Clark 2018); we group the other two options into the "neutral" class. Specifically, we develop a framework that accurately assesses whether a given premise entails a given hypothesis using background knowledge provided by external sources including WordNet (Miller 1995), ConceptNet (Speer, Chin, and Havasi 2017), and DBPedia (Auer et al. 2007).

Usable inference and reasoning methods have been a central contribution of artificial intelligence research. Specifically, reasoning methods and the knowledge bases that support them have played an important role in addressing formal reasoning tasks. Our own previous work (Boratko et al. 2018) associates the QA task with the reasoning and knowledge types required to answer various types of science exam questions. While the introduction of large datasets for NLI has led to it being cast as a classification problem, the potential of pre-existing knowledge has been only minimally explored (Chen et al. 2018; Marelli et al. 2014). This knowledge may manifest itself in various forms: as facts consisting of entities and their relationships; as lexical knowledge about the synonyms of entities; or as common-sense concepts and relationships between them. Exploiting additional knowledge from different knowledge graphs may well aid attempts to solve the NLI problem.

Figure 1 shows an example of a subgraph that can be derived from ConceptNet based on the concepts mentioned in a given premise and hypothesis[1]. This subgraph includes connections between the concepts mentioned in the premise and the hypothesis, via concepts available in ConceptNet. Such subgraphs, enhances the concepts mentioned in the text with additional information and hence can improve the performance of learning-based systems (Kapanipathi et al. 2014;

[1]Example based on the SNLI dataset (Bowman et al. 2015).

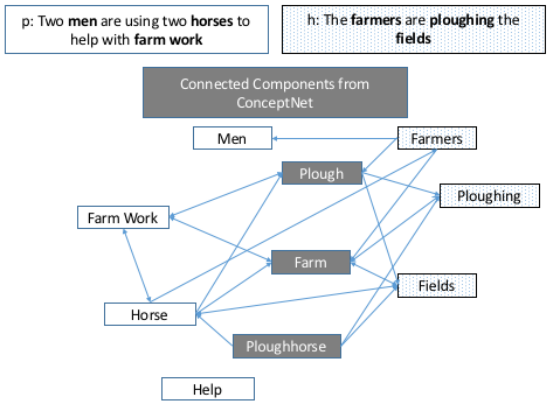

Figure 1: Example of a subgraph from ConceptNet for a given premise `p` and a hypothesis `h`. Edges represent the existence of a relationship between concepts; nodes (concepts) with a gray background are those not mentioned explicitly in either `p` or `h`.

Bouchoucha, He, and Nie 2013), particularly for NLI.

Contributions. In this paper, we introduce the **`ConSeqNet`** framework, which enables the use of various kinds of external knowledge bases to retrieve knowledge relevant to a given NLI instance, by retrieving information related to the premise and hypothesis. We describe our novel architecture and demonstrate its use with a specific external knowledge source – ConceptNet – and evaluate its performance on two other sources, WordNet and DBpedia. We compare the performance of three distinct approaches to augmenting the knowledge used to train for and to predict entailment relationships between given pairs of premises and hypotheses: graph-only, text-only, and text-and-graph. Using both qualitative and quantitative results, we demonstrate that introducing graph-based features boosts performance on the NLI problem, but only when text features are present as well. Our system has a competitive performance (accuracy) of 85.2% on the SciTail entailment dataset, which is derived from science domain question answering datasets.

2 Background and Related Work

We consider prior work in the areas of both the Natural Language Inference, and Knowledge Graphs[2]; as well as work at their intersection.

2.1 Natural Language Inference

Recently, the NLI task has gained significant attention due to the release of multiple crowd-sourced, large-scale datasets that can be used to train neural network classifiers (Parikh et al. 2016; Zhao, Huang, and Ma 2016; Khot, Sabharwal, and Clark 2018; Chen et al. 2018; Yin, Roth, and Schütze 2018). In particular, the SNLI (Bowman et al. 2015) dataset has been a major catalyst of research on NLI facilitating learning-based approaches by providing a sufficient number of training examples for data-intensive learning algorithms. The Multi Genre NLI corpus (MultiNLI) (Williams, Nangia, and Bowman 2018) is an effort to address the limitations of SNLI. The dataset focuses on domain adaptation by introducing genre labels for each sentence pair. SciTail (Khot, Sabharwal, and Clark 2018), derived from science domain multiple choice question datasets (Welbl, Liu, and Gardner 2017; Clark et al. 2016), has the focus on the down-stream task of standardized-test question-answering.

Neural networks with an encoder-attention-classifier architecture are commonly used for NLI tasks (Bowman et al. 2015; Wang, Hamza, and Florian 2017; Khot, Sabharwal, and Clark 2018). To encode the premise and hypothesis, most methods use RNNs, which are used for many NLP tasks. Bowman et al. (2015) used LSTMs to learn sentence representations for premise and hypothesis separately, concatenating them for classification. Rocktäschel et al. (2015) introduced a word-by-word attention model that learns conditional encodings of premise and hypothesis for textual entailment. The match-LSTM model (Wang and Jiang 2015) extended the model presented by Rocktäschel et al. (2015) to address the limitation of a single vector representation of the premise for learning attention weights. The next step in this direction was the multi-perspective matching mechanism introduced by Wang, Hamza, and Florian (2017). Most of these text-based models differ from each other in their attention mechanisms. While the above-mentioned approaches used word-by-word attention mechanisms, Yin, Roth, and Schütze (2018) has explored inter-sentence interactions for textual entailment on the SciTail dataset; however, other NLI datasets were not used for evaluation. Recently, Glockner, Shwartz, and Goldberg (2018) exposed the simplicity of the existing neural network models by showing that they do not perform well on a new test-set that has significantly lesser overlap between the premise and hypothesis. The best performing model on this test set is Chen et al. (2018) which harnesses WordNet as external knowledge; providing added motivation to explore external knowledge sources for NLI.

2.2 External Knowledge (Knowledge Graphs)

Most openly available external knowledge sources are in the form of graphs, commonly known as Knowledge Graphs (KG). Some of the well known KGs include Freebase, DBpedia, Yago, and ConceptNet. A KG is defined as a set of concepts connected by relationships where the concepts form the nodes of the graph and the relationships are the labeled edges. A fact such as "Barack Obama is the spouse of Michelle Obama" is represented in a KG (for example on DBpedia) as `dbr:Barack_Obama dbo:spouse dbr:Michelle_Obama`, where `dbr:Barack_Obama` and `dbr:Michelle_Obama` are nodes and `dbo:spouse` is a labeled edge.

KGs have been used extensively in Information Retrieval (Bouchoucha, Liu, and Nie 2014), Recommendation Systems (Lalithsena et al. 2017; Kapanipathi et al. 2014) and Question Answering (Sun et al. 2018). The prominent concern when using KGs is the availability of relevant knowledge, both in terms of applicability to the task setting and

[2]We use External Knowledge and Knowledge Graphs interchangeably in this paper.

to the domain/topic of interest. This is closely related to the way in which these knowledge graphs are created. For instance, DBpedia is a generic knowledge base with information extracted from Wikipedia infoboxes; these contain facts such as (*Barack_Obama, spouseof, Michelle_Obama*). By contrast, ConceptNet (Liu and Singh 2004) consists of common-sense knowledge acquired through crowd sourcing. While DBpedia is well suited for entity-based tasks such as movie recommendation and entity disambiguation, ConceptNet may be more appropriate if common sense reasoning is required. WordNet is a lexical database offering synonyms, hypernyms, hyponyms, and antonyms. In this work, we experiment with DBpedia, ConceptNet, and WordNet.

2.3 Knowledge Graphs and NLI

Few prior approaches have exploited syntactic structures in the form of graphs for textual entailment (Zhao, Huang, and Ma 2016; Khot, Sabharwal, and Clark 2018). However, the graph structures used in these models do not come from external knowledge sources and are not used to enhance the textual content of either the premise or hypothesis. One work which closely relates to the objective of the paper, i.e., enhancing the model with external knowledge is Chen et al. (2018). This work harnesses WordNet as the external knowledge for NLI. WordNet, however, is a lexical database restricted to a small number of linguistic relationships among terms. Furthermore, Chen et al. (2018) uses only four relationships to generate five features based on WordNet. In our work, including WordNet, we explore more expressive KGs such as DBpedia and ConceptNet. Also, the most important difference is the way we use KGs: our method enriches the texts (as graphs) and builds a matching model over the enriched texts for textual entailment. Another interesting alternative direction for using KGs such as WordNet for NLI is presented by Kang et al. (2018). The approach generates adversarial examples for robust training of NLI systems, but this improves performance only during limited supervision.

3 Approach

In this section, we present the architecture of our system called **`ConSeqNet`**, and the approach underlying it. Figure 2 depicts the framework of **`ConSeqNet`** – the system can use both textual as well as structured information from KGs to assist in determining textual entailment. The framework can be clearly divided into two parts: (a) a **text based model** that takes in as input the premise and hypothesis text; and (2) a **graph based model** whose input is specific knowledge derived from the knowledge base using the given premise and hypothesis.

3.1 Text Based Model

Most models developed for NLI have relied on the text of the given premise and hypothesis, without the aid of any external knowledge. These models follow an encode-attend-classify framework. First, the premise and hypothesis are encoded using recurrent neural networks (RNNs). Next, an attention layer is implemented on top of the encoders. The final layer is then used for classification. In our work, we have primarily used match-LSTM (Wang and Jiang 2015) as our text based model. We opted for match-LSTM for two main reasons: (1) match-LSTM as an entailment model has been proved useful for multiple NLP tasks (Wang and Jiang 2015; Wang et al. 2017) and (2) our implementation of match-LSTM performed significantly better than the baselines (details in Section 4).

Given a premise P (with K words) and a hypothesis H (with J words), the model computes the matching results between them as follows:

- **Context Encoding:** A contextual representation of the premise and hypothesis is generated by first transforming premise P and hypothesis H into their embedding vectors $\mathbf{t_i^p}$ and $\mathbf{t_j^h}$, where $\mathbf{t_i^p}$ and $\mathbf{t_j^h}$ are embedding vectors of i-th word in premise and j-th word in hypothesis. These embedding vectors are further encoded using BiLSTMs to generate the context encodings of premise and hypothesis. Let $\mathbf{p}_i$ and $\mathbf{h}_j$ be the contextual representation of the i-th word of premise and the j-th word of hypothesis.
- **Word-by-Word Attention:** This layer computes the inter-attention between the contextual embeddings of the premise and hypothesis. The entries of the (unnormalized) attention matrix $\mathbf{E} \in \mathbb{R}^{K \times J}$ are defined as:

$$E_{ij} = \mathbf{p}_i \cdot \mathbf{h}_j \tag{1}$$

  we can then compute the soft alignment α_j for the hypothesis as follows:

$$\alpha_j = \sum_{i=1}^{K} \frac{\exp(E_{ij})}{\sum_{k=1}^{K} \exp(E_{kj})} \mathbf{p}_i \tag{2}$$

- **Matcher:** We compare the soft aligned premise and hypothesis at each word position as a feature vector:

$$\tilde{\mathbf{h}}_j = [\mathbf{h}_j; \alpha_j; \mathbf{h}_j - \alpha_j; \mathbf{h}_j \odot \alpha_j],$$

  where $\odot$ denotes the element-wise multiplication, and $[;]$ denotes vector concatenation. The feature vectors for all the hypothesis positions are fed into a BiLSTM to get the matching states $\{\mathbf{h}_j^{\mathrm{m}}\}_{j=1:J} = \mathrm{BiLSTM}(\{\tilde{\mathbf{h}}_j\}_{j=1:J})$.
- **Pooling:** To get a fixed-sized vector representation for the matching result, we apply max-pooling:

$$\mathbf{x}_{\mathrm{out}}^{\mathrm{text}} = \max([\mathbf{h}_1^{\mathrm{m}}, \mathbf{h}_2^{\mathrm{m}}, \ldots, \mathbf{h}_J^{\mathrm{m}}]) \tag{3}$$

  $\mathbf{x}_{\mathrm{out}}^{\mathrm{text}}$ is then used for classification.

3.2 Graph Based Model

The input to the graph based model is a premise graph and a hypothesis graph respectively. As shown in Figure 2, the first step is to transform the given premise and hypothesis text into a relevant subgraph mapped to an external knowledge. Given premise text P and hypothesis text H, the transformation generates $P_g = (V_p, E_p)$ and $H_g = (V_h, E_h)$ where $V_p = (v_1^p, ..., v_K^p)$ and $V_h = (v_1^h, ..., v_J^h)$ are a subset of concepts in the knowledge graph. E_p and E_h are labeled edges connecting concepts in P_g and H_g respectively. H_g and P_g are the input for our graph-based entailment model.

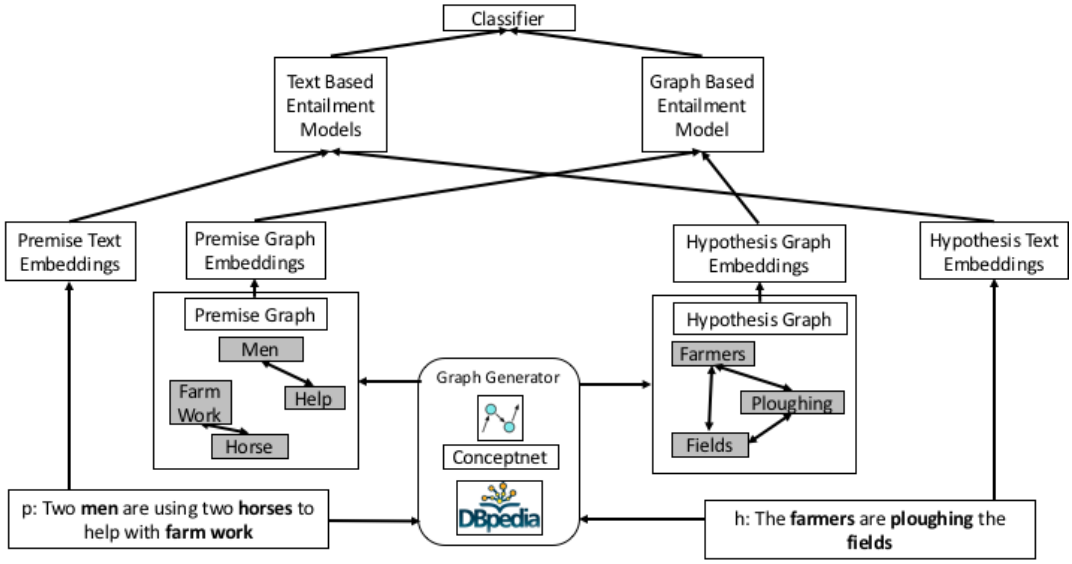

Figure 2: The overall architecture of the **ConSeqNet** system, illustrated via an example.

Premise and Hypothesis Graphs Premise and hypothesis graphs are generated by mapping text phrases to concepts in the external knowledge graph. For instance, as shown in Figure 1, given hypothesis *"The farmers are ploughing the fields"*, we map "*farmers*", "*ploughing*, and "*fields*" in the sentence to their corresponding concepts in the external knowledge. We use the following techniques to generate three different types of subgraphs for each premise and each hypothesis separately:

- *Concepts Only*[3]*:* For each premise and hypothesis, we use lexical mappings within each knowledge source to map individual words and phrases to concepts. Those concepts make up the vertices of the *Concepts Only* subgraph. All edges connecting those vertices are also included in the subgraph.
- *One-Hop:* The concepts in the *Concepts Only* graph for premise and hypothesis are expanded to include all their one hop neighbors in the external knowledge source.
- *Two-Hop:* Two-Hop graphs were generated by adding two hop neighbors to the *Concepts Only* graph, but only if they constitute paths between vertices that were already in the *Concepts Only* graph. This can be thought of as a graph that increases interconnectedness by introducing minimal new concepts, hence reducing noise. We took this approach because we observed that *One-Hop* graphs turned to be noisy (on average, the *One-Hop* graphs generated using ConceptNet increased the number of concepts extracted from premise text from 9 to 326).

We use the graphs constructed (under the three conditions above) for premise and hypothesis as the input to our neural network model for entailment. In this work, we explore two different neural network models: (1) Gmatch-LSTM; and (2) Graph concepts attention model (GconAttn).

Gmatch-LSTM for Concepts Only Graphs Our first design choice was to use match-LSTM with the *Concepts Only* graphs P_g and H_g. The key idea is to treat each node (concept) in the vertex set of a given graph (V_p or V_h) as a single token, and re-arrange them into a sequence of concepts. This sequence is ordered by the length of the concept and its positions in the original texts. Thus, match-LSTM can be applied to match the sequence of premise concepts and the sequence of hypothesis concepts.

In contrast with the text based model, where each word is first represented as its word embedding vector, in the graph based match-LSTM each concept is initialized using the corresponding *Concept Embedding*. The concept embeddings are trained by knowledge graph embedding techniques such as TransH (Wang et al. 2014), and CompleX (Trouillon et al. 2016) using the corresponding knowledge graph.

GconAttn Model for General Graphs The use of match-LSTM is justified when the input is the *Concepts Only* graph, since the concept can be aligned in the order of their appearance in premise and hypothesis to form text sequences. However, it is non-trivial to order the concepts in the graphs generated by the *One-Hop* and *Two-Hop* strategies; this makes the use of match-LSTM unintuitive. We therefore develop a new model *GconAttn* (Graph Concepts Attention Model) to overcome this limitation.

The embeddings of premise and hypothesis concepts from the graph are the input to the neural network. An attention weighted representation of both premise and hypothesis is

[3]Concepts in the *Concepts Only* graph are used to form the base set for *One-Hop* and *Two-Hop* graph generation techniques.

determined, and is transformed to a final fixed size representation using pooling. This representation of premise and hypothesis (concatenated) is used for classification. Below, we distinguish the graph based model from the match-LSTM (text based) model described previously:

- **Context Encoding:** Since the concept graphs may not have sequential structures, the GconAttn model directly starts with the **concept embeddings**.
- **Word-by-Word Attention:** This layer computes the inter-attention between the embeddings of the concepts in premise and hypothesis to find the best aligned concepts between the respective graphs. The major difference from match-LSTM is that GconAttn performs two-way attention (Seo et al. 2016). The model computes the soft alignment for premise and hypothesis respectively as follows:

$$\beta_i = \sum_{j=1}^{J} \frac{\exp(E_{ij})}{\sum_{k=1}^{J}\exp(E_{ik})}\mathbf{h}_j, \quad \alpha_j = \sum_{i=1}^{K} \frac{\exp(E_{ij})}{\sum_{k=1}^{K}\exp(E_{kj})}\mathbf{p}_i$$

- **Matcher:** Due to the lack of sequential structure, the matching nodes of each hypothesis concept j or premise concept i are computed with a projection feed-forward network:

$$\mathbf{p_i^m} = F\left([\mathbf{p}_i; \beta_i; \mathbf{p}_i - \beta_i; \mathbf{p}_i \odot \beta_i]\right)$$
$$\mathbf{h_j^m} = F\left([\mathbf{h}_j; \alpha_j; \mathbf{h}_j - \alpha_j; \mathbf{h}_j \odot \alpha_j]\right)$$

  where $F(\cdot)$ denotes a feed-forward network; $\odot$ denotes the element-wise multiplication; and $[;]$ denotes vector concatenation.

- **Pooling:** Finally, we apply max-pooling and mean-pooling across all matching nodes in the premise and hypothesis. The output of pooling is concatenated, and is presented as the output of the graph model, which can then be used to make the final prediction. $\mathbf{s} \in \{\mathbf{h}, \mathbf{p}\}$.

$$\mathbf{s}'_{\text{max}} = \max\left([\mathbf{s_1^m}, \mathbf{s_2^m}, \ldots, \mathbf{s_N^m}]\right) \quad (4)$$
$$\mathbf{s}'_{\text{avg}} = \text{avg}\left([\mathbf{s_1^m}, \mathbf{s_2^m}, \ldots, \mathbf{s_N^m}]\right)$$
$$\mathbf{x}_{\text{out}}^{\text{graph}} = \left[\mathbf{p}'_{\text{max}}; \mathbf{p}'_{\text{avg}}; \mathbf{h}'_{\text{max}}; \mathbf{h}'_{\text{avg}}\right]$$

3.3 Merging Text and Graph Models

The final results of matching the text model $\mathbf{x}_{\text{out}}^{\text{text}}$ and the graph model $\mathbf{x}_{\text{out}}^{\text{graph}}$ are concatenated, and passed on to a feed forward network that classifies between *entailment* and *neutral*, which are the two classes in the SciTail dataset:

$$\mathbf{pred} = F\left(\left[\mathbf{x}_{\text{out}}^{\text{text}}; \mathbf{x}_{\text{out}}^{\text{graph}}\right]\right) \quad (5)$$

It is important to note that the final hidden state of the graph model can be directly fed into a classifier. This can be considered an entailment model that uses only information from external knowledge.

4 Experiments and Results

In this section, we detail the experiments that we perform to evaluate our **ConSeqNet** system. We first describe each dataset, followed by the training setup of our best model and implementation details. We then compare that model's performance to numbers from recent entailment models and baselines. In the latter parts of this section, we discuss the selection of knowledge graphs and the performance of graph construction techniques – specifically, the *Concepts Only, One-Hop,* and *Two-Hop* methods (c.f. Section 3).

4.1 Datasets

We use the SciTail dataset (Khot, Sabharwal, and Clark 2018), which is a textual entailment dataset derived from publicly released science domain multiple choice question answering datasets (Welbl, Liu, and Gardner 2017; Clark et al. 2016). The dataset contains 27,026 sentence pairs (premise and hypothesis), with binary labels denoting whether the relationship between each pair is `entails` or `neutral`. The hypothesis is created using the question and correct answer from the options; the premise is retrieved from the ARC corpus (data.allenai.org/arc/arc-corpus. The performance of the entailment models on this dataset is shown in Table 1; the evaluation metric is accuracy.

Model	Dev	Test
Decomp-Attn (Parikh et al. 2016)	75.4	72.3
DGEM* (Khot, Sabharwal, and Clark 2018)	79.6	77.3
DeIsTe (Yin, Roth, and Schütze 2018)	82.4	82.1
BiLSTM-Maxout (Mihaylov et al. 2018)	-	84.0
match-LSTM (Wang and Jiang 2015)	88.2	84.1
Our implementation		
match-LSTM (GRU)	88.5	84.2
match-LSTM+WordNet* (Chen et al. 2018)	88.8	84.3
match-LSTM+Gmatch-LSTM* (**ConSeqNet**)	**89.6**	**85.2**

Table 1: Performance of entailment models on SciTail in comparison to our best model that uses match-LSTM as the text and the graph model with *Concepts Only* graph and CN-PPMI embeddings. * indicates the use of external knowledge in the approach.

4.2 Training Setup

All words in the text model are initialized by 300D Glove vectors (Glove 840B 300D) (nlp.stanford.edu/projects/glove), and the concepts that act as the input for the graph model are initialized by 300D ConceptNet PPMI vectors (Speer, Chin, and Havasi 2017); these are openly available for ConceptNet. We use the pre-trained embeddings without any fine tuning. We have adapted match-LSTM with GRUs as our text and graph based model. The system is trained by Adagraph with a learning rate of 0.001, and batch size of 40. Both the text and graph based models are trained jointly. For the graph model, we use concepts from the *Concepts Only* graph, which is generated using the approach detailed in Section 3.2.

4.3 Implementation

We used the AllenNLP (allennlp.org) library to implement all the models used in the experiments. While the previous section (Training Setup) specifically focused on our best

performing model, here we provide implementation details for all the experiments that were performed. We used a plug-and-play approach where we varied: (a) text models (match-LSTM, DeCompAttn); (b) graph models (Gmatch-LSTM, GconAttn); (c) external knowledge sources (DBpedia, WordNet, and ConceptNet); (d) graph construction techniques (*Concepts Only, One Hop,* and *Two Hop*); and (e) graph embeddings (CN-PPMI, TransH).

In order to map text to concepts, we used DBpedia Spotlight (Mendes et al. 2011) for DBpedia, and Spacy (spacy.io) to implement a max-substring match for WordNet and ConceptNet. While ConceptNet had openly available embeddings (CC-PPMI), we used TransH embeddings (Wang et al. 2014) generated using OpenKE (openke.thunlp.org) with default configurations for the other knowledge sources.

4.4 Baselines and Comparison

We compare our work to the following baselines: (1) Decomposable Attention Model (Decomp-Att) (Parikh et al. 2016); (2) Decomposed Graph Entailment Model (DGEM) (Khot, Sabharwal, and Clark 2018), which is the first of the entailment models to use graph structure from OpenIE as external knowledge and show improvements on SciTail; (3) Deep explorations of Inter-sentence interactions for Textual entailment (DeIsTe) (Yin, Roth, and Schütze 2018); (4) BiLSTM-Maxout (Mihaylov et al. 2018), the latest entailment model that has shown promise on the SciTail dataset[4]; (4) match-LSTM (Wang and Jiang 2015), which has shown good performance on the SNLI dataset; (5) match-LSTM with GRU, which replaces LSTMs with GRUs for its encoding, since GRUs give better empirical results (this is also the match-LSTM model used in our **ConSeqNet** model); and (5) match-LSTM – WordNet features Chen et al. (2018), which uses five features based on synonyms, antonyms, hypernyms, and co-hypernyms from WordNet (external knowledge) in its co-attention mechanism to improve performance on SNLI. We reimplement these five features and add them to baseline (5).

Table 1 shows the performance of our **ConSeqNet** model in comparison to other entailment models. Two of the five models in Table 1 use some kind of external knowledge (indicated by *). Almost all the match-LSTM variations exhibit accuracy greater than 84%, which is better than the recent entailment models on which the SciTail dataset has been tested. Similar to the results shown in (Yin et al. 2017), using GRUs with match-LSTM slightly improved the accuracy (particularly on the dev set). The accuracy of our jointly trained model **ConSeqNet** is 89.6% on the dev set and 85.2% on the test set. Later in this paper, we expound further on our model – this includes choices such as the use of different knowledge sources, graph models, and the graph construction technique.

[4]The details of the BiLSTM-Maxout (Mihaylov et al. 2018) model are not available. We have reported the numbers provided in a pre-print of a publication that is to appear.

	DBpedia	WordNet	ConceptNet
Entities/Concepts	5M	155K	1.1M
Relationships	1100	16	40
Facts	33M	117K	3.15M

Table 2: Comparison between different knowledge sources based on the number of entities, relationships, and facts.

4.5 Selecting External Knowledge Source

In this work, we focus on openly available knowledge graphs. Based on the knowledge graphs' availability for use and their distinct properties, we chose DBpedia, ConceptNet, and WordNet. In Table 2, we provide details on each of these knowledge graphs. DBpedia is the largest with more than 5 million entities and 33 million facts. ConceptNet subsumes WordNet conceptually, and both contain general type information; however, the reliability of WordNet's linguistic features is higher than ConceptNet's.

For the NLI problem, WordNet has only been partially explored. WordNet is a lexical database with a restricted set of relationships between terms. Chen et al. (2018) use WordNet by creating five new knowledge-based features that are added (for co-attention) to their model. The five features are derived from synonym/antonym and hypernym/hyponym relationships in WordNet between terms in the premise and hypothesis. In Table 1, we show that using the expanded set of features from WordNet – used in the attention mechanism of match-LSTM – has an accuracy of 84.3% on the SciTail datset. While WordNet is useful, it is restrictive in terms of its applicability, particularly because of its coverage and the types of relationships available. DBpedia and ConceptNet are larger knowledge sources, as shown in Table 2.

In terms of the number of concepts mapped from text to external knowledge, ConceptNet has more concepts (9) on average in comparison to DBpedia (6). While it is important that the graphs are not noisy, it is also important to have enough information to exploit. In which case, ConceptNet is slightly better with few more concepts per sentence.

The types of relationships between concepts in DBpedia are factual and derived from Wikipedia. Based on qualitative analysis of these relationships, we found that they may not be suitable for NLI datasets, and specifically SciTail. On the other hand, we found that, ConceptNet with its 40 relationships expressing common sense knowledge may be more suitable. In order to quantitatively select the right knowledge source, we determined the impact of the each knowledge graph on SciTail. We created the *Concepts Only* graph from each of the available external knowledge sources, and evaluated them using our match-LSTM+GconAttn model on the dev set of SciTail. The results of this experiment are shown in Table 3; based on these results, we decided to use ConceptNet for all our graph-based experiments (detailed next).

4.6 Graph Generation Experiments

In order to select the graph generation mechanism (c.f. Section 3.2) that has the highest impact on the NLI problem, we ran experiments with match-LSTM as our text model and GconAttn as our graph model. GconAttn was chosen be-

Knowledge Sources	Accuracy
WordNet	87.6
DBpedia	87.3
ConceptNet	**88.6**

Table 3: Results using match-LSTM+GconAttn with different knowledge sources on SciTail dev set.

cause the concepts retrieved from *One-Hop* and *Two-Hop* do not have any specific sequence or ordering. Table 4 presents the results of these experiments with the same hyperparameters for all three graph generation experimental conditions.

All the graph+text models perform equally well. However, when only the graph model is considered, the *One-Hop* graph exhibits lower accuracy in comparison to the *Concepts Only* and *Two-Hop* graphs. This may be due to the noise induced from external knowledge and the addition of a large number of concepts in the *One-Hop* case. On average, premise and hypothesis sentences consist of 19 and 12 words respectively, but their respective *One-Hop* graphs have over 300 concepts. On the other hand, the *Concepts Only* graphs average 9 and 6 concepts with better performance than the *One-Hop* graphs (72.3% vs 68.2%). Based on these results, and the simplicity of the graph construction technique, we pursued further experiments with the *Concepts Only* graphs.

The accuracy of the graph only models are relatively low, whereas the graph+text and text only models are comparable. This might lead to the conclusion that the graph+text model is only driven by the text. However, an *Oracle* model condition, where we choose the correct answer between the text and graph models, indicates that the graph model contributes to improved accuracy on the dev set (88.5% for text only from Table 1 versus at least 91.6% for Oracle). With *One-Hop*, the graph model correctly predicts almost 40% of the answers that are incorrectly predicted by the text-only model. We thus conclude that there is value to using external knowledge for NLI on SciTail. This is contrary to Mihaylov et al. (2018)'s conclusion that external knowledge from ConceptNet is not useful in this domain.

Graph Generation	Graph Model Accuracy	Graph + Text Model Accuracy	Oracle Text ∨ Graph	Avg Concepts Premise (19 words)	Avg Concepts Hypothesis (12 words)
Concepts	72.3	87.2	**92.5**	9	6
One Hop	68.2	87.3	**93.1**	369	312
Two Hop	71.7	87.3	**91.6**	23	15

Table 4: Performance of graph generation techniques on the match-LSTM+GconAttn model on the SciTail dev set.

4.7 Selecting Text+Graph Model

We experimented with many text models and selected match-LSTM due to its superior performance on SciTail. In order to determine the best combination of the graph models (Gmatch-LSTM and GconAttn) with match-LSTM as the text model, we ran multiple experiments on the SciTail dev set. The results are shown in Table 5. Both GconAttn and Gmatch-LSTM are competitive in their performance. We also experimented with different knowledge graph embeddings to determine their impact on these models. ConceptNet-CN-PPMI clearly shows an improvement **(89.6)** in comparison to other models on the dev set. This model is used as our final model to be evaluated and compared against the baselines, leading to new state of the art performance numbers as shown in Table 1.

Text Model	Graph Model	Embedding	Dev
match-LSTM	Gmatch-LSTM	ConceptNet-CN-PPMI	**89.6**
match-LSTM	Gmatch-LSTM	ConceptNet-TransH	87.3
match-LSTM	Gmatch-LSTM	ConceptNet-CompleX	88.3
match-LSTM	GconAttn	ConceptNet-CN-PPMI	88.6
match-LSTM	GconAttn	ConceptNet-TransH	87.6
match-LSTM	GconAttn	ConceptNet-CompleX	88.4

Table 5: Results of combinations of text and graph models along with various ways of computing embeddings on Scitail dev set.

5 Conclusion & Future Work

In this paper, we presented the **`ConSeqNet`** system: an entailment model for solving the Natural Language Inference (NLI) problem that utilizes ConceptNet as an external knowledge source. Our model provides performance that is close to state-of-the-art, with an accuracy of 85.2% on the SciTail dataset. We analyze various external knowledge sources and their effect on NLI, and show – in direct contrast to other recent studies – that there is promise in using knowledge graphs such as ConceptNet for textual entailment.

Our future work includes designing a framework to exploit multiple relevant knowledge sources based on the given dataset and context. Existing external knowledge sources are known to be extremely noisy, and new techniques must be developed in order to extract knowledge relevant to a specific task (such as NLI). Another interesting direction involves exploring new ways to represent the structure of premise and hypothesis subgraphs, and systematically using the relations between the concepts contained therein to improve performance on the NLI task.

References

Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*. 722–735.

Boratko, M.; Padigela, H.; Mikkilineni, D.; Yuvraj, P.; Das, R.; McCallum, A.; Chang, M.; Fokoue-Nkoutche, A.; Kapanipathi, P.; Mattei, N.; Musa, R.; Talamadupula, K.; and Witbrock, M. 2018. A Systematic Classification of Knowledge, Reasoning, and Context within the ARC Dataset. In *Machine Reading for Question Answering (MRQA) Workshop at*.

Bouchoucha, A.; He, J.; and Nie, J.-Y. 2013. Diversified query expansion using conceptnet. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, 1861–1864. ACM.

Bouchoucha, A.; Liu, X.; and Nie, J.-Y. 2014. Integrating multiple resources for diversified query expansion. In de Rijke, M.; Kenter, T.; de Vries, A. P.; Zhai, C.; de Jong, F.; Radinsky, K.; and Hofmann, K., eds., *Advances in Information Retrieval*, 437–442.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Chen, Q.; Zhu, X.; Ling, Z.-H.; Inkpen, D.; and Wei, S. 2018. Neural natural language inference models enhanced with external knowledge. In *Proceedings of ACL 2018*.

Clark, P.; Etzioni, O.; Khot, T.; Sabharwal, A.; Tafjord, O.; Turney, P. D.; and Khashabi, D. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *30th*, 2580–2586.

Glockner, M.; Shwartz, V.; and Goldberg, Y. 2018. Breaking NLI systems with sentences that require simple lexical inferences. *arXiv preprint arXiv:1805.02266*.

Kang, D.; Khot, T.; Sabharwal, A.; and Hovy, E. 2018. Adventure: Adversarial training for textual entailment with knowledge-guided examples. *arXiv preprint arXiv:1805.04680*.

Kapanipathi, P.; Jain, P.; Venkataramani, C.; and Sheth, A. 2014. User interests identification on twitter using a hierarchical knowledge base. In *European Semantic Web Conference*, 99–113.

Khot, T.; Sabharwal, A.; and Clark, P. 2018. SciTail: A textual entailment dataset from science question answering. In *32nd*.

Lalithsena, S.; Perera, S.; Kapanipathi, P.; and Sheth, A. 2017. Domain-specific hierarchical subgraph extraction: A recommendation use case. In *Big Data (Big Data), 2017 IEEE International Conference on*, 666–675.

Liu, H., and Singh, P. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal* 22(4):211–226.

MacCartney, B., and Manning, C. D. 2009. *Natural language inference*. Stanford University Stanford.

Marelli, M.; Bentivogli, L.; Baroni, M.; Bernardi, R.; Menini, S.; and Zamparelli, R. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, 1–8.

Mendes, P. N.; Jakob, M.; García-Silva, A.; and Bizer, C. 2011. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, 1–8.

Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answerin. In *2018*.

Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Parikh, A. P.; Täckström, O.; Das, D.; and Uszkoreit, J. 2016. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*.

Rocktäschel, T.; Grefenstette, E.; Hermann, K. M.; Kočiskỳ, T.; and Blunsom, P. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

Seo, M.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, 4444–4451.

Sun, H.; Dhingra, B.; Zaheer, M.; Mazaitis, K.; Salakhutdinov, R.; and Cohen, W. 2018. Open domain question answering using early fusion of knowledge bases and text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4231–4242.

Trouillon, T.; Welbl, J.; Riedel, S.; Gaussier, É.; and Bouchard, G. 2016. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, 2071–2080.

Wang, S., and Jiang, J. 2015. Learning natural language inference with lstm. *arXiv preprint arXiv:1512.08849*.

Wang, Z.; Zhang, J.; Feng, J.; and Chen, Z. 2014. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, 1112–1119.

Wang, S.; Yu, M.; Jiang, J.; Zhang, W.; Guo, X.; Chang, S.; Wang, Z.; Klinger, T.; Tesauro, G.; and Campbell, M. 2017. Evidence aggregation for answer re-ranking in open-domain question answering. *arXiv preprint arXiv:1711.05116*.

Wang, Z.; Hamza, W.; and Florian, R. 2017. Bilateral multi-perspective matching for natural language sentences. *arXiv preprint arXiv:1702.03814*.

Welbl, J.; Liu, N. F.; and Gardner, M. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text at*, 94–106.

Williams, A.; Nangia, N.; and Bowman, S. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of NAACL: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122. Association for Computational Linguistics.

Yin, W.; Kann, K.; Yu, M.; and Schütze, H. 2017. Comparative study of cnn and rnn for natural language processing. *CoRR* abs/1702.01923.

Yin, W.; Roth, D.; and Schütze, H. 2018. End-task oriented textual entailment via deep explorations of inter-sentence interactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, 540–545.

Zhao, K.; Huang, L.; and Ma, M. 2016. Textual entailment with structured attentions and composition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2248–2258.