Knowledge-Based Systems 219 (2021) 106889

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Frame-based Neural Network for Machine Reading Comprehension

Shaoru Guo [a,*], Yong Guan [a], Hongye Tan [a,b], Ru Li [a,b,*], Xiaoli Li [c]

[a] School of Computer & Information Technology, Shanxi University, China
[b] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, China
[c] Institute for Infocomm Research, A*Star, Singapore

ARTICLE INFO

Article history:
Received 22 September 2020
Received in revised form 18 February 2021
Accepted 20 February 2021
Available online 23 February 2021

Keywords:
Machine reading comprehension
Frame semantics
Frame representation
Multiple-Frame semantic integration

ABSTRACT

Machine Reading Comprehension (MRC) is one of the most challenging tasks in Natural Language Understanding (NLU). In particular, MRC systems typically answer a question by only utilizing the information contained in a given piece of text passage itself, while human beings can easily understand the meanings of the passage based on their background knowledge. To bridge the gap, we propose a novel *Frame-based Neural Network for Machine Reading Comprehension*(FNN-MRC) method, which employs Frame semantic knowledge to facilitate question answering. Specifically, different from existing Frame based methods that only model lexical units (LUs), our FNN-MRC has a Frame representation model, which utilizes both LUs in Frame and Frame-to-Frame (F-to-F) relations, designed to model Frames and sentences (in passage) together with attention schema. In addition, FNN-MRC has a Frame-based Sentence Representation (FSR) model, which is able to integrate multiple-Frame semantic information to obtain much better sentence representation. As such, FNN-MRC explicitly leverages the above Frame knowledge to assist its semantic understanding and representation. Extensive experiments demonstrate that our FNN-MRC method is able to achieve better results than existing state-of-the-art techniques across multiple datasets.

1. Introduction

Machine Reading Comprehension (MRC) requires machines to read and understand text passages, and answer relevant questions about it. It is regarded as an effective way to measure language understanding and typically requires a deep understanding of the given passage in order to answer its question correctly. Clearly, human beings can easily understand the meanings of a text passage based on their background knowledge. For instance, given a sentence *Katie bought some chocolate cookies*, people know *Katie* is a *buyer*, and *chocolate cookies* are *goods* that belong to *Food* class etc. Existing machine learning approaches, however, face great challenges to address the complicated MRC questions, as they do not have above semantic background knowledge.

Nevertheless, FrameNet [1,2], as a knowledge base, provides schematic scenario representation that could be potentially leveraged to facilitate text understanding. It enables the development of wide-coverage Frame parsers [3,4], as well as various real-world applications, ranging form event recognition [5], textual entailment [6], question answering [7], narrative schemas [8] and paraphrase identification [9], etc. In particular, a *Frame* (F) is defined as a composition of *Lexical Units* (LUs) and a set of *Frame Elements* (FEs). Given a sentence, if its certain word/phrase evokes a Frame by matching a LU, then it is called *Target* (T). It is worth mentioning that FrameNet arranges different relevant Frames into a network by defining Frame-to-Frame (F-to-F) relations. Fig. 1 provides an example of F, FEs, LUs, T and F-to-F, where the target word **bought** in sentence *Katie bought some chocolate cookies* evokes a Frame ***Commerce_buy*** as it matches with a LU **buy** (*bought* is the past tense of *buy*). In addition, another target word **chocolate cookies** evokes a different Frame ***Food***. Finally, a couple of relevant Frames, including ***Commerce_buy***, ***Shopping***, ***Seeking***, ***Locating***, form **F-to-F** relations.

There exist other semantic resources, such as WordNet [10], PropBank [11]. In particular, WordNet is a lexicon that clusters words into sets of synonyms (synsets) and describes semantic relationships between them. In comparison, FrameNet annotates sentences/examples with both syntactic and semantic information for each Lexical Unit, which clearly provides more rich information than WordNet. On the other hand, PropBank is a corpus annotated with argument role labels for verbs. Roles in PropBank are general, while Roles in FrameNet are specific to lexical unit (Buyer vs. Arg0, Goods vs. Arg1).

The code (and data) in this article has been certified as Reproducible by Code Ocean:https://help.codeocean.com/en/articles/1120151-code-ocean-s-verification-process-for-computational-reproducibility. More information on the Reproducibility Badge Initiative is available at www.elsevier.com/locate/knosys.

* Corresponding authors at: School of Computer & Information Technology, Shanxi University, China.
E-mail addresses: guoshaoru0928@163.com (S. Guo), guanyong0130@163.com (Y. Guan), tanhongye@sxu.edu.cn (H. Tan), liru@sxu.edu.cn (R. Li), xlli@i2r.a-star.edu.sg (X. Li).

https://doi.org/10.1016/j.knosys.2021.106889
0950-7051/

FrameNet	Frame (F)	**Commerce_buy**	
		Definition: These are words describing a basic commercial transaction involving a Buyer and a Seller exchanging Money and Goods, taking the perspective of the Buyer.	
	Frame Elements (FEs)	**Buyer**	The Buyer wants the Goods and offers Money to a Seller in exchange for them.
		Goods	The Goods is anything (including labor or time, for example) which is exchanged for Money in a transaction.
		Seller	The Seller has possession of the Goods and exchanges them for Money from a Buyer
	Lexical Unites (LUs)	**buy.v**, buy.n, buyer.n, client.n, purchase.v, purchaser.n, ...	
	Target (T)	[Katie]$_{Buyer}$ **bought**$_{Commerce\_buy}$ [some chocolate cookies]$_{Goods}$.	
	Frame-to-Frame Relation (F-to-F)	Commerce buy—Shopping—Seeking—Locating	
WordNet	**Synset (buy.v)**: buy, purchase; **Synset (buy.n)**: bargain, steal		
PropBank	[Katie]$_{Arg0}$ **bought** [some chocolate cookies]$_{Arg1}$.		

Fig. 1. Example of F, FEs, LUs, T and F-to-F in FrameNet, and the comparison with other two semantic resources WordNet and PropBank.

It is clear that FrameNet can bring in additional background semantic knowledge that could be leveraged to improve MRC performance. However, how to effectively utilize these useful semantic knowledge from FrameNet is an important issue. Previously, feature-based [12] supervised learning models were proposed to integrate Frame knowledge to MRC, where they require language experts design complex features, which is typically a time consuming and expensive process and may not be generic enough to handle different MRC tasks. Later, end-to-end solutions with neural models [13–15] achieved good performance on the MRC tasks. Although such techniques can effectively incorporate contextual information from large-scale external unlabeled data into machine learning models, we still lack of effective representation learning techniques to help us incorporate Frame knowledge into a good representation so that we can leverage to build a successful MRC system. In addition, we observe the existing works mainly focus on LU embedding within a Frame [16–18], without *modeling a Frame as a whole*. Furthermore, many sentences could have more than one target words that evoke multiple semantically correlated Frames, but existing methods do not focus on integrating multi-Frame from FrameNet to enrich accurate and comprehensive sentence semantic representations.

To address the problems mentioned above, in this paper, we propose a novel *Frame-based Sentence Representation* (FSR) model, which leverages rich Frame semantic knowledge, including both generalizations of LUs and F-to-F relations, to better model the sentences in given text passage. To take full advantages of LUs and F-to-F, we propose three different strategies for Frame representation. Finally, we integrate multiple-Frame semantic information to get more comprehensive sentence representation based on individual Frame representation.

In this paper, we propose a Frame-based Neural Network for MRC (FNN-MRC). Specifically, we first utilize the FSR model to capture the multiple Frame semantic information of every sentence, and GRU [19] is used to aggregate a document-level frame-based representation. In experiments, we evaluate our FNN-MRC method on multi-choice MRC task, such as MCTest [20], *non-extractive* MRC, which requests to choose the right option from a set of candidate answers according to given passage and question. This is different from relatively easy extractive MRC datasets such as SQuAD [21] and NewsQA [22], which require a model to extract an answer span to a question from reference passage. In *non-extractive* MRC, however, machine learning models need to perform reasoning and inference. In addition, its difficulty is also reflected by the required background knowledge that are not expressed in given passage. We show improvements on two widely used neural models, i.e., traditional deep learning methods (with LSTM [23]) and Transformer (with BERT [15]) in the experiments.

The key contributions of this work can be summarized as follows:

1. We propose novel attention-based *Frame Representation Models*, which take full advantage of LUs and F-to-F relations to model Frames with attention schema.
2. We propose a new *Frame-based Sentence Representation* (FSR) model that integrates multi-Frame semantic information to obtain richer semantic aggregation for better sentence representation.
3. We propose a *Frame-based Neural Network for MRC* (FNN-MRC), which explicitly leverages the Frame representation and Frame-based sentence representation knowledge to assist non-extractive question answering.
4. Our experimental results demonstrate our *Frame-based Neural Network for MRC* (FNN-MRC) is very effective on Machine Reading Comprehension (MRC) task, comparing with state-of-the-art techniques.

2. Related work

In this section, we will first provide a brief introduction to MRC datasets, which play an important role in recent progress in reading comprehension, and subsequently describe machine learning models applied specifically to two MRC datasets, namely, MCTest and RACE, in details.

2.1. Publicly available datasets for reading comprehension tasks

Machine reading comprehension has become one of central tasks in natural language understanding, fueled by the creation of many large-scale datasets. Existing datasets on MRC are categorized into three groups: *cloze-style*, *question answering* and *multiple-choice reading comprehension*.

Cloze-style reading comprehension task is to infer the missing entity in the query by understanding the content of given article. Hermann et al. [13] published the CNN and Daily Mail, which paired news articles with their summarized bullet points. Cui et al. [24] firstly published the Chinese cloze-style reading comprehension dataset, which consisted of People Daily news dataset and Children's Fairy Tale (CFT) dataset. Cloze-style shares most of the characteristics of reading comprehension, but the answer is a single word in the article [25]. So these datasets rarely test the deep reasoning capabilities of the underlying machine learning models.

Question answering reading comprehension task requires models to form a span in the passage to answer the question. Rajpurkar et al. [21] proposed SQuAD, the answer to each question was always a span in the context. In addition, Joshi et al. [26] proposed TriviaQA, which only needed to form the span that seems most related to the question, instead of checking whether the answer is actually entailed by the passage. Since annotators tend to copy spans as answers directly, the majority questions in these datasets are still extractive answers.

For multiple-choice MRC datasets, given a question and a text passage, we need to select the correct answer from multiple candidate answers. This is a non-extractive task, in which answers are not restricted to extractive text spans and we should infer the answer based on the semantics of passage and question. In particular, MCTest [20] was built by crowdsourcing, which involved extensive human effort in designing questions and answers. More recently, RACE [27] dataset has been proposed where its passages and questions was complied by human experts (English instructors), which were carefully designed for evaluating the middle and high school Chinese students' ability in language understanding and logic reasoning. Note this multiple-choice MRC task is quite different from previous two groups of datasets discussed above — besides surface matching between candidate answers to given passage, it focuses more on semantic interpretation, summarization and the use of prior background knowledge. So in this paper, we focus on more challenging multiple-choice MRC datasets MCTest and RACE.

2.2. Machine learning methods

Machine Learning Methods be categorized into *classical machine learning methods* and *neural methods*. Next, we elaborate them in details.

Classical Machine Learning Methods aim at mapping a sequence of texts (given passage, questions and candidate answers) into a feature representation based on semantic or syntactic features to select the best candidate answer according to given passage and question. Simple text matching methods, such as window-based and distance-based algorithms, only use lexical features to count the number of overlapping words among the given passage, question, and answers [20]. Smith et al. [28] further improved the lexical matching method by taking into account multiple context windows, question types and coreference resolution. As the simple text matching methods could not handle complex questions, more expressive *hidden variable models* are thus introduced. For instance, Wang et al. [12] used a simple latent-variable classifier trained with a max-margin criterion by augmenting baseline features [20] with additional features based on syntax, Frame semantics, coreference, and word embeddings. Narasimhan and Barzilay [29] proposed a discriminative Framework with hidden variables that focused on discourse relation features. Sachan and Xing [30] and Sachan et al. [31] proposed unified max-margin Framework that utilized hidden variables to find the alignment of question, correct answer, and passage. Sachan and Xing [30] proposed a Framework that learned to find the latent mapping of the question–answer meaning representation graph onto the passage meaning representation graph using the Abstract Meaning Representation (AMR) formalism. Sachan et al. [31] presented a Framework that learned to find the hidden structure that explained the relation between the question, correct answer, and passage based on rhetorical structure, coreference links. Lu et al. [32] proposed multilayer network embedding algorithms based on Nonnegative Matrix Factorization (NMF), which can characterize the nature of a node with different structural property constraints. Li et al. [33] studied interaction among hidden variables for the machine comprehension by using linguistic structures to help capturing global evidence in hidden variable modeling.

Features in these methods often require significant effort to design, and rely on various NLP tools to extract, which are typically time consuming and erroneous, and hard to apply in other tasks/applications.

Neural models first encode given passage and concatenation of question and one candidate answer (do four concatenations as we usually have four candidate answers) into two vectors and then compute the similarity between them. The candidate answer with largest similarity will be chosen as the correct answer. After the release of large-scale MRC datasets, various neural models have been proposed recently. They built the representation of words and then fed them into a deep neural network which processes and compares the presentations between passage and question + answer, where attention mechanism is mainly used to model their interactions. For example, Lai et al. [27] adopted and modified Gated Attention Reader (GA) [34] and Stanford Attention Reader (Stanford AR) [35] for both cloze-style MRC and multi-choice MRC problems. However, experimental results showed that these models are not capable of tackling the tasks effectively. Yin et al. [36] explored a hierarchical attention-based convolutional neural network (HABCNN) for multi-choice MRC, which employed an attention mechanism to detect key phrases, key sentences and key snippets that are relevant to the question. Xu et al. [37] proposed the Dynamic Fusion Networks (DFN), which used multi-hop reasoning mechanism and employed reinforcement learning techniques for dynamic strategy selection for this task. Tay [38] proposed Multi-Range Reasoning Units (MRU), a new compositional encoder which constructed gates from a novel contract-and-expand operation for multiple choice MRC. Wang et al. [39] proposed a new co-matching approach to this problem, which jointly modeled whether a passage can match both a question and a candidate answer. Wen et al. [40] proposed a multilabel image classification method, which aims at projecting both labels and image features to a common latent vector space. In this way, the frequently occurring labels and features do appear closer in the latent space.

Research [15,41,42] demonstrated that by fine-tuning a *pre-trained language model* can lead to a series of breakthroughs in MRC. Among the pre-trained language models, BERT [15] has taken the MRC world by storm. So many optimized versions of BERT have been proposed. For instance, SpanBert [43] extended BERT by masking contiguous random spans, rather than random tokens. RoBERTa [44] used a novel dataset for pre-training to improve performance on downstream tasks. Zhang et al. [45] proposed a dual co-matching network (DCMN), which modeled the relationship among passage, question and answers bidirectionally based on BERT. Megatron-BERT [46] implemented a simple and efficient model parallel approach using intra-layer model-parallelism.

It is worth noting we use BERT as the backbone to illustrate how our proposed method works for its superior performance in a range of natural language understanding datasets, although many other pre-trained optimized models can be applied as well.

Neural models + External Knowledge. As external knowledge plays a critical role in MRC, some existing work have explored different ways to leverage external knowledge. For instance, some work leverage *data augmentation* to improve MRC performance. For instance, Pan [47] improved question answering by utilizing intra-domain external question answering datasets and enriching the reference corpus by out-domain external corpora. Wang et al. [48] built an attention-based recurrent neural network model with the help of external knowledge, which was semantically relevant to the current machine comprehension task.

Note pre-trained language models have leveraged rich knowledge by pre-training deep neural models with language model objectives over large-scale unlabeled corpora (e.g., Wikipedia articles). For example, Sue et al. [49] improved machine reading comprehension task during fine-tuning stage, instead of incorporating more prior knowledge into a model via pre-training. More specifically, they fine-tune a pre-trained language model with reading strategies identified in cognitive science on the multiple-choice MRC task.

In this paper, we aim to utilize Frame semantic knowledge to improve multiple-choice question answering during the fine-tuning stage. Our method explicitly leverages LUs and Frame-to-Frame relations to model Frames and effectively integrates multi-Frame semantic information to obtain richer semantic information.

3. Frame representation model

For each sentence in given passage, we can get its corresponding Frame semantic annotations by Frame Annotator SEMAFOR [4], which will clearly bring additional background semantic knowledge to help us to better tackle MRC task. In this section, we present our *Frame representation model* to represent the semantic information of Frames, considering Lexical Units (LUs), Frame-to-Frame (F-to-F) relations and corresponding sentence. In particular, Frame (F) is defined as a composition of LUs and different relevant Frames that are arranged into a network based on F-to-F relation. As shown in Fig. 2, we can find that the lexical units (LUs) of Frame ***Commerce_buy*** contain *buy.n, buy.v, buyer.n, client.n, purchaser.n, purchase.v* et al. In addition, Frame ***Seeking***, ***Shopping***, ***Commerce_buy***, ***Locating***, ***Scrutiny*** have relevant semantic relations according to the relations defined in FrameNet. For example, Frame ***Shopping*** inherits from Frame ***Seeking***.

Let $F = \{F_1, F_2, \ldots, F_m, \ldots\}$ represents a set of all Frames in FrameNet, where $F_m \in \mathcal{R}^H$ is the representation of m-th Frame of F. Let $U^{F_m} = \{u_1^{F_m}, u_2^{F_m}, \ldots, u_n^{F_m}, \ldots\}$ be the LU set of a Frame F_m, where $U^{F_m} \in \mathcal{R}^{(H \cdot N)}$, N stands for the total number of LUs in F_m and H is the frame dimension, and $u_n^{F_m}$ be the n-th LU of F_m. t^{F_m} is a target word, matching a LU in F_m. We propose 3 different Frame representation models.

3.1. Lexical Unit Aggregation Model (LUA)

Lexical Unit Aggregation Model (LUA) is a straightforward idea. As shown in Fig. 3, given a Frame F_m, it averages all its underlying LU representation $u_n^{F_m}$ ($u_n^{F_m} \in U^{F_m}$) to represent the overall Frame:

$$F_m = \frac{1}{N} \sum_{U^{F_m}} u_n^{F_m} \tag{1}$$

For Frame ***Commerce_buy***, we first map all the LUs (i.e., Buy, Client, Purchase, Purchaser) to their corresponding vector representations, and subsequently average their representations to finally obtain the representation of Frame ***Commerce_buy***.

3.2. Lexical Unit Attention Model (TLUA)

Each Frame in above LUA model has the same representation for different sentences, as they do not distinguish the importance of each LU in the Frame with regards to its corresponding sentence from given passage. To address this issue, we propose a Lexical Unit Attention Model (TLUA) model, utilizing an attention scheme to automatically weight different LUs for the Frame, according to target word T in the corresponding sentence, shown in Fig. 4.

More specifically, we compute the weighted sum of target word T's representation and other LUs' representations based on their importance w.r.t. T. In other words, we emphasize T as it occurs in the context of given sentence where only certain LUs should play more important roles in Frame representation. This TLUA can thus reduce the potential noise introduced by irrelevant LUs in the same Frame by taking semantics of given sentence into consideration. For example, in sentence *She bought some chocolate cookies*, *bought* evokes the Frame ***Commerce_buy***. In this case, **bought** called target word (T), which is more important than other LUs (such as *client*) under the Frame ***Commerce_buy***. It should be noted that we encode multiple word target by averaging of all word representations in it.

$$F_m = t^{F_m} + \sum_{\widetilde{U}^{F_m}} att(u_n^{F_m}) \cdot u_n^{F_m} \tag{2}$$

$$att(u_n^{F_m}) = \frac{exp(t^{F_m} \cdot u_n^{F_m})}{\sum_{u_k^{F_m} \in \widetilde{U}^{F_m}} exp(t^{F_m} \cdot u_k^{F_m})} \tag{3}$$

Here, $\widetilde{U}^{F_m}$ represents the LU set of F_m which does not include t^{F_m}, and $\widetilde{U}^{F_m} \in \mathcal{R}^{H \cdot (N-1)}$.

3.3. Frame Relation Attention Model (FRA)

We propose a novel Frame Relation Attention Model (FRA), which takes advantage of F-to-F relations to get much richer semantic information without introducing too much noise, shown in Fig. 5.

Given Frame F_m, $F_m^+ = \{F_{m,1}, \ldots, F_{m,w}, \ldots\}$ represents its expanded Frames, including all the Frames that can be linked to F_m through F-to-F relation chains in FrameNet. Note attention schemes have been designed for both *intra-Frame* and *inter-Frames*. Particularly, intra-Frame attention focuses on relevant LUs, while inter-Frames attention emphasizes relevant Frames, avoiding the influence from less relevant but linked Frames. For example, given the sentence *She bought some chocolate cookies*, *intra-Frame* pays more attention to *bought* and *inter-Frames* pays more attention to Frame ***Shopping*** (instead of Frame ***Scrutiny***), when modeling the Frame ***Commerce_buy***.

$$F_m^* = F_m + \sum_{w=1}^{W} att(F_{m,w}) \cdot F_{m,w} \tag{4}$$

$$att(F_{m,w}) = \frac{exp(F_m \cdot F_{m,w})}{\sum_{k=1}^{W} exp(F_m \cdot F_{m,k})} \tag{5}$$

4. Frame-based sentence representation

Given a sentence $s = \{x_1, x_2, \ldots, x_k, \ldots\}$, where each x_k is a word, let T_k be the k-th Frame-evoking target of s, and T_k evokes F_k Frame. FE_{ki} denotes the i-th *Frame element* of F_k, and P_{ki} denotes the i-th span fulfilling FE_{ki}. We define a Frame semantic quadruple $c_k = \langle T_k, F_k, FE_{kn}, P_{kn} \rangle$, where c_k represents the k-th quadruple of s.

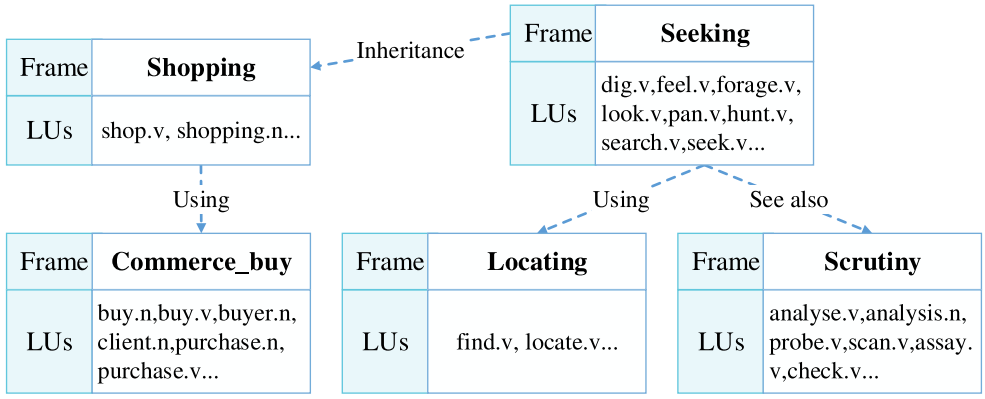

Fig. 2. An example of Frame-to-Frame relations according to the lexical units.

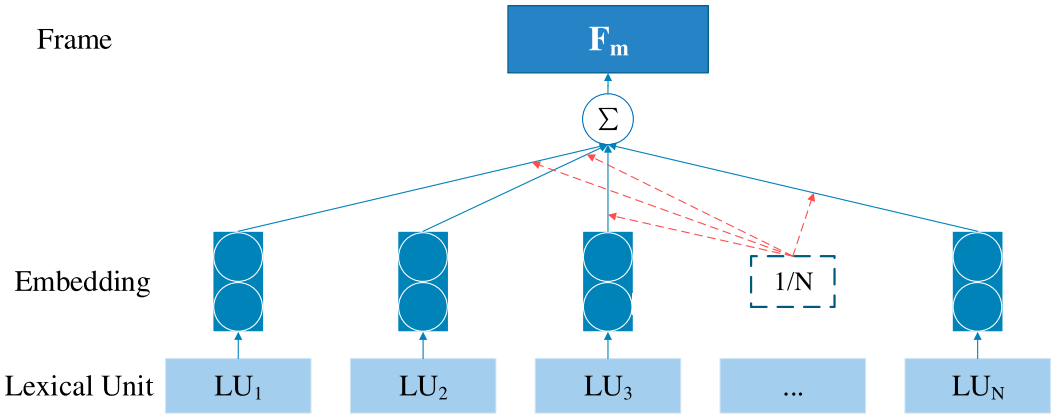

Fig. 3. Lexical Unit Aggregation Model.

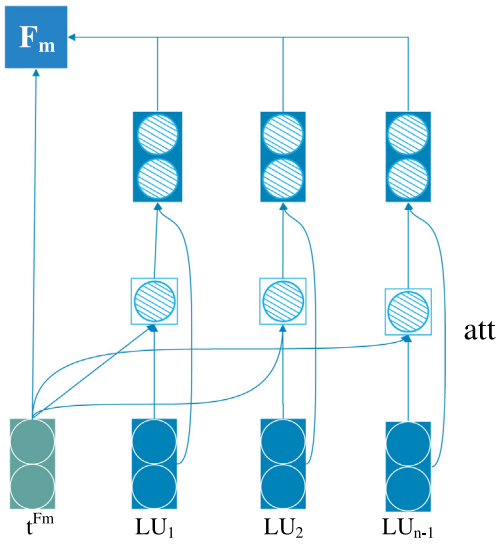

Fig. 4. Lexical Units Attention Model.

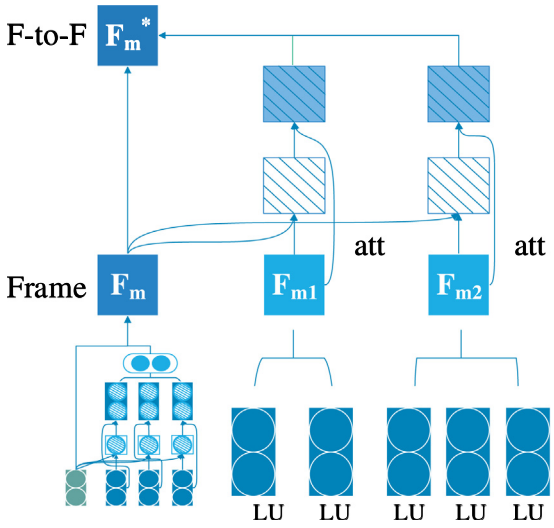

Fig. 5. Frame Relation Attention Model.

4.1. Sentence semantic annotations with multiple frames

In this paper, we employ SEMAFOR [4] to automatically process sentences and assign them with multiple semantic annotations [50].

Fig. 6 provides an example sentence with three T, namely *bought*, *some*, *chocolate cookies*. Each T has its evoked semantic Frame right below it. For each Frame, its FEs are shown enclosed in the block where dark gray indicates the corresponding T, and the words fulfilling the FEs are connected to the corresponding text. For example, T **bought** evokes the ***Commerce_buy*** Frame, and has the *Buyer*, *Goods* FEs fulfilled by *Katie* and *some chocolate cookies*.

The sentence s in Fig. 6 has three quadruples:

1. $c_1 = \langle T_1, F_1, FE_{1n}, P_{1n} \rangle$ =<bought, *Commerce_buy*, [Buyer, Goods], [Katie, chocolate cookies]>

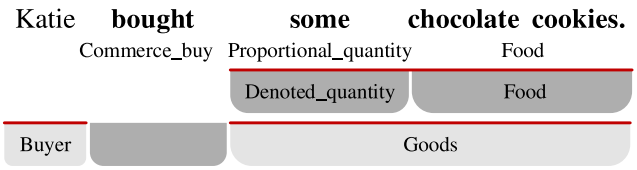

Fig. 6. A Sentence of FrameNet Annotations.

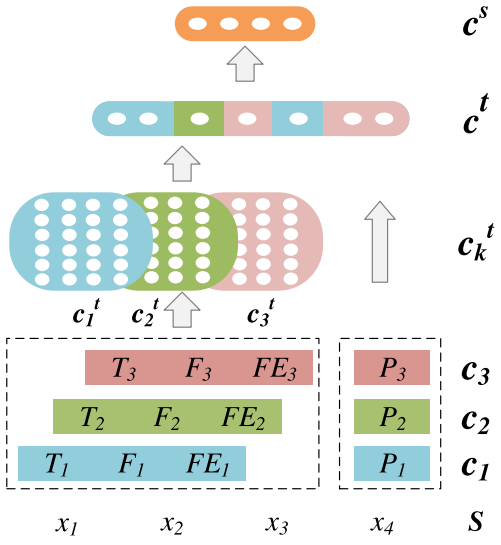

Fig. 7. Frame Integration Representation Model.

2. $c_2 = \langle T_2, F_2, FE_{2n}, P_{2n}\rangle$ =< some, *Proportional_quantity*, [Denoted_quantity], [some]>

3. $c_3 = \langle T_3, F_3, FE_{3n}, P_{3n}\rangle$ =<chocolate cookies, *Food*, [Food], [chocolate cookies]>

4.2. Frame Integration Representation (FIR)

In Fig. 7, c_k (k=1, 2, 3) is the input. We first compute its representation c_k^t, with columns denoting different semantic information. Target embeddings are pretrained by GloVe [51]. FEs embeddings are initialized randomly with a uniform distribution. We use three different methods, namely, LUA, TLUA, TFA, for Frame representation according to the lexical unit embeddings, and we emphasize T as it occurs in the given sentence.

Then, we formalize sentence representation as follows:

$$c^s = \mathcal{N}(c^t) \tag{6}$$

$$c^t = \phi(c_k^t) \quad (k = 1, \ldots K) \tag{7}$$

Where K represents the total number of quadruples in the sentence. $\phi(c_k^t)$ is aggregate operation, used to form Frame set representation c^t based on the information of P and T in the sequence. Finally, we encode sentence information by neural network models $\mathcal{N}(\cdot)$.

Next, we introduce aggregate operation $\phi(c_k^t)$. Given a sentence, we first compute the semantic coverage of every Frame $cov(F_i^s)$, which is simply based on the ratio of the number of Frame annotation words $num(F_i)$ to the total number of words in the source sequence $num(s)$.

$$cov(F_i^s) = \frac{num(F_i)}{num(s)} \tag{8}$$

For instance, the coverage of *Commerce_buy* cov(*Commerce_buy*) in sentence [*Katie*]$_{Buyer}$ **bought**$_{Commerce\_buy}$ [*somechocolate cookies*]$_{Goods}$ is 1, as the total number of words in the source sequence $num(F_i)$ and the number of Frame annotation words num(*Commerce_buy*) are both 5. While the coverage of *Food* cov(*Food*) in sentence *Katie bought some* [**chocolatecookies**]$_{Food}$ is 0.4, as the number of Frame annotation words num(*Food*) is 2.

We then iteratively replace words with Frames and Frame elements embedding according to the coverage of every Frame, based on the information of P and T position information in the source sequence.

5. Frame-based Neural Network for Machine Reading Comprehension

Frame-based Neural Network for Machine Reading Comprehension (FNN-MRC) architecture comprises three key components: raw context representation, Frame-based context representation and answer prediction. The architecture is illustrated in Fig. 8. In this section, we provide the details on how to implement the three components and explain how they work.

5.1. Task definition

The multiple-choice machine comprehension task can be formulated as follows: $\langle P, Q, A\rangle$, where Q is the question, $A = \{a_1, a_2, \ldots, a_i, \ldots, a_N\}$ is a candidate answer set for the question, and $N = 4$ in this paper (as well as in many datasets), which means there are four candidate answers for every question and we need to choose one correct answer from A. $P = \{s_1, s_2, \ldots, s_j, \ldots, s_M\}$ is the text passage, and s_j represents the j-th sentence of P. The objective of machine reading comprehension is to select a best answer a^* ($a^* \in A$) for question Q according to the passage P.

5.2. Raw context representation

We construct the input as: the passage as sequence A, and the concatenation of question and one candidate answer as sequence B. The raw context can be denoted as: *RawContext* = [[CLS] Passage [SEP] Question + Candidate Answer[SEP]]. The single input sequence *RawContext* is then fed into Neural Network to get its representation c^r.

$$c^r = NN(RawContext) \tag{9}$$

We consider two widely used neural models: (i) traditional deep learning methods (with LSTM [23]), (ii) ubiquitous transformer architecture [52] (with BERT [15]).

When the neural network is Bi-LSTM, we first run LSTM on *RawContext* independently, and then aggregate their vectorized representations into a vector.

$$\overrightarrow{c^r} = \overrightarrow{BiLSTM}(RawContext) \tag{10}$$

$$\overleftarrow{c^r} = \overleftarrow{BiLSTM}(RawContext) \tag{11}$$

$$c^r = [\overrightarrow{c^r}; \overleftarrow{c^r}] \tag{12}$$

Specifically, Bi-LSTM consists of a forward network and a backward network, where the forward network $\overrightarrow{BiLSTM}(\cdot)$ handles *RawContext* from left to right, and the backward network $\overleftarrow{BiLSTM}(\cdot)$ processes it in reverse order. Here, $\overrightarrow{c^r}$ is the forward hidden state from the forward network and $\overleftarrow{c^r}$ is the backward hidden state from the backward network. Finally, we concatenate $\overrightarrow{c^r}$ and $\overleftarrow{c^r}$, resulting in c^r.

Alternatively, if BERT is employed to encode context, then we formulate it as follows:

$$c^r = BERT(RawContext) \tag{13}$$

Where c^r is the contextual representation using the BERT encoder.

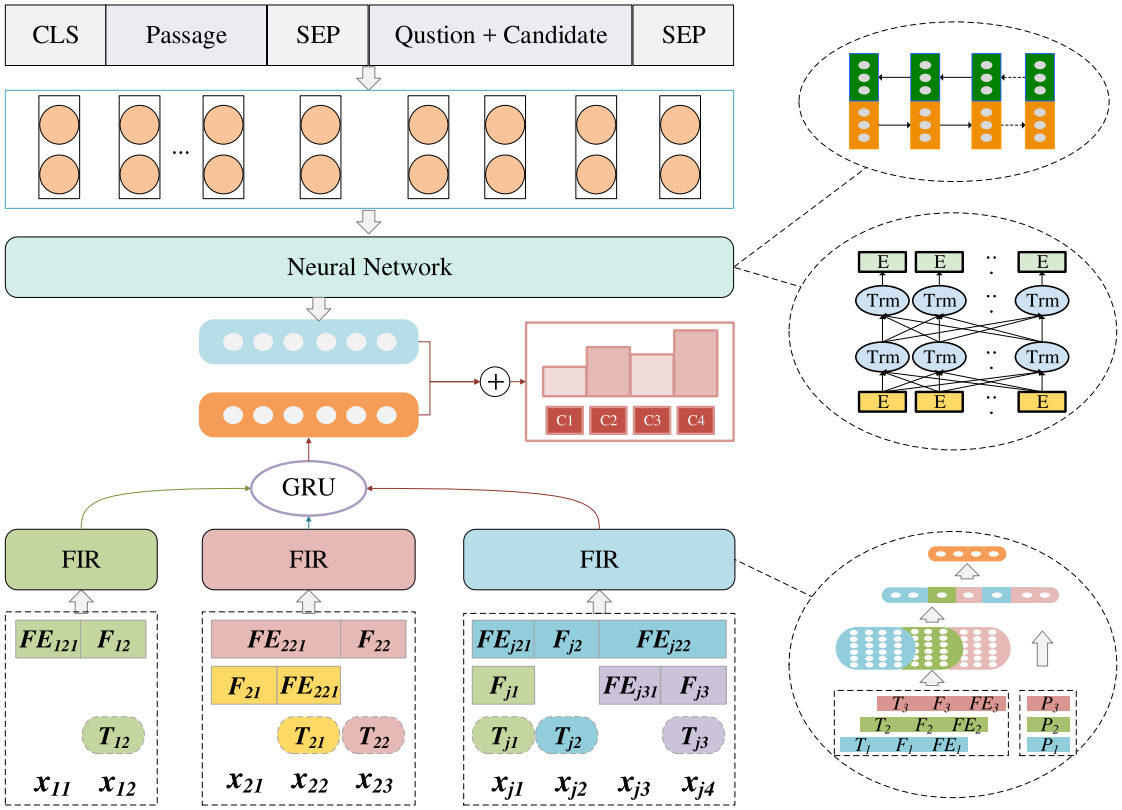

Fig. 8. Frame-based Neural Network for Machine Reading Comprehension.

5.3. Frame-based context representation

Frame-based context representation aims at distilling semantic information from text. Given a passage $P = \{s_1, s_2, \ldots, s_j, \ldots, s_M\}$, question Q and a candidate answer a_i, we utilize the FIR model, described in 4.2, to capture the multiple Frame semantic information of every sentence. Note we will get $M + 2$ Frame-based sentence representation $c^s = [c_1^s, c_2^s, \ldots, c_M^s, c_{M+1}^s, c_{M+2}^s]$, that is, M sentences of passage, 1 question sentence and 1 candidate answer sentence.

After the Frame semantics of every sentences are sufficiently modeled, GRU [19] is used to aggregate a document-level representation c^f. We define the function as:

$$c^f = GRU(c_1^s, c_2^s, \ldots, c_M^s, c_{M+1}^s, c_{M+2}^s) \tag{14}$$

Where GRU (Gated Recurrent Unit) is a type of recurrent neural networks. We feed Frame-based sentence representation c^s into GRU, and regard the output of the GRU c^f as the Frame-based document representation.

5.4. Answer prediction

After we obtain the raw context representation c^r from 5.2 and the Frame-based context representation c^f from 5.3 separately, a concatenate function is used to merge both two representations to get the final context representation c^{rf}.

$$c^{rf} = concat(c^r; c^f) \tag{15}$$

Where $concat(\cdot)$ represents concatenation operation, which is used to combine the hidden states of the c^r and c^f. And c^{rf} is the final representation, containing context and Frame semantic information.

Finally, an answer prediction is drawn from the softmax distribution over the scores of the four candidate answers. Specially, we apply a linear layer and a softmax layer on the final hidden state c^{rf} and the model predicts the answer with the maximal probability across all the candidate answers:

$$c_i = softmax(g(c^{rf})) \tag{16}$$

$$a^* = argmax(c_i) \tag{17}$$

Where $g(\cdot)$ is a linear function, and $softmax(\cdot)$ is used to normalize the final hidden state c^{rf} to a probability distribution c_i. a^* is the correct answer, which has the maximum value $argmax(c_i)$ among candidate answers.

6. Experiments

In this section, we conduct comprehensive experiments to compare our FNN-MRC model with existing state-of-the-art techniques. To better analyze the performance of our FNN-MRC method on MRC, we consider two types of neural models: (i) traditional deep learning methods LSTM [23], (ii) the powerful pre-trained language model. For pre-trained model, we use BERT as the backbone to illustrate how the proposed method works, as its superior performance in a range of MRC tasks.

6.1. Datasets for MRC

For experiments, we employ two benchmark datasets, namely, MCTest [20] and RACE [27], to evaluate the system performance of multiple-choice machine comprehension task.

In particular, **MCTest** [20] consists of two subsets, namely MCTest-160 and MCTest-500. They are from open-domain, yet restricted to concepts and words that a seven-year-old child is expected to understand. In addition, it contains a corpus of fictional story sets that were constructed with a crowd-sourcing method.

Table 1
The performance comparison of different models on two MCTest.

Method	MCTest-160 (%)	MCTest-500 (%)
Richardson et al. [20]	69.16	63.33
Wang et al. [12]	75.27	69.94
Li et al. [33]	74.58	72.67
Attentive Reader [13]	46.3	41.9
Neural Reasoner [53]	47.6	45.6
Reading Strategies [49]	81.7	82.0
Bert [45]	73.8	80.4
BERT+DCMN+ [45]	85.0	86.5
FNN-MRC	86.1	84.2

Table 2
Performance comparison with three different frame representation models on MCTest.

Method	MCTest-160 (%)	MCTest-500 (%)
Bert [45]	73.8	80.4
Bert (Our implementation)	82.5	80.9
Bert+LUA	82.7	79.5
Bert+TLUA	84.6	82.7
Bert+FRA (FNN-MRC)	86.1	84.2
bi-LSTM	54.2	49.5
bi-LSTM+LUA	59.4	57.5
bi-LSTM+TLUA	61.5	58.2
bi-LSTM+FRA (FNN-MRC)	62.7	59.6

RACE [27] also consists of two subsets: RACE-M and RACE-H. RACE-M comes from middle school English examinations while RACE-H comes from high school English examinations in China, both of which were constructed from real-world examinations.

In total, we have four datasets from **MCTest** and **RACE** data.

6.2. Existing models

We compare our model with a number of baseline models. Now we briefly introduce several representative models.

Sliding Window [20,27] computes the matching score based on the matched words between the question–answer pair and passage with a fixed window size.

Co-Match [39] treats the question and the candidate answer as two sequences and jointly models whether a passage can match both a question and a candidate answer.

Reading Strategies [49] aims at improving machine reading comprehension by fine-tuning a pre-trained language with three reading strategies (i.e., back and forth reading, highlighting, and self-assessment).

BERT [15]) inputs the passage as sentence A and the concatenation of the question and the candidate answer as sentence B, and applies a softmax layer for selecting the answer.

BERT+DCMN+ [45] uses dual co-matching network (DCMN) to model the relationship among passage, question and answer options bidirectionally, and also adds *passage sentence selection* and *answer option interaction* into the model to select the answer.

6.3. Experiment results on MCTest

Table 1 shows our FNN-MRC model achieves 86.1% accuracy on MCTest-160, which is significantly better than all the eight state-of-the-art methods. In addition, it also achieves very competitive results on MCTest-500, i,e, much better than seven existing methods, slightly worse than BERT+DCMN + model. This is encouraging, as our model is much simpler than BERT+DCMN+, which uses much more sophisticated architecture and is hard to transfer to other tasks.

Table 3
The performance comparison of different models on RACE.

Method	RACE-M (%)	RACE-H (%)	RACE (%)
Sliding Window [27]	37.3	30.4	32.2
Stanford AR [35]	44.2	43.0	43.3
Co-Match [39]	55.8	48.2	50.4
OpenAI GPT [54]	62.9	57.4	59.0
Reading Strategies [49]	69.2	61.5	63.8
BERT+DCMN+ [45]	79.2	72.1	74.1
bi-LSTM	53.5	45.3	47.7
FNN-MRC(bi-LSTM)	57.2	47.4	50.3
Bert [45]	76.6	70.1	72.0
FNN-MRC(Bert)	81.3	73.5	75.8
RoBERTa [55]	86.5	81.3	83.2
Amazon Mechanical Turker	85.1	69.4	73.3
Human Ceiling Performance	95.4	94.2	94.5

Table 4
Performance comparison with three different frame representation models on RACE.

Method	RACE-M (%)	RACE-H (%)	RACE (%)
Bert [45]	76.6	70.1	72.0
Bert+LUA	79.8	71.6	74.1
Bert+TLUA	80.7	72.8	75.2
Bert+FRA (FNN-MRC)	81.3	73.5	75.8
bi-LSTM	53.5	45.3	47.7
bi-LSTM+LUA	54.9	45.8	48.5
bi-LSTM+TLUA	56.1	46.3	49.2
bi-LSTM+FRA (FNN-MRC)	57.2	47.4	50.3

Recall in Section 3, we proposed three different methods, namely, LUA, TLUA, FRA, for Frame representation. Table 2 shows their detailed results. We have the following three observations:

(1) No matter for BERT or Bi-LSTM, if we add Frame semantic information, the performance improves by several percents, indicating Frame information is very valuable in helping semantic understanding.

(2) Comparing TLUA with LUA, TLUA performs better, signifying attention scheme in TLUA can capture semantic information more accurately.

(3) Finally, FRA further improves LUA and TLUA's performance, as sentences within a passage typically have semantic connections with each other, and it is thus necessary to take advantage of F-to-F relations to enrich semantic information.

6.4. Experiment results on RACE

We further conduct experiments on RACE (RACE-M and RACE-H) and the results are shown in Table 3. We report the performance of the following models: majority baselines, our FNN-MRC method with two widely used encoding models (BERT and bi-LSTM), optimized versions of BERT-style model, and human performance. Note Turkers is the performance of Amazon Turkers on a randomly sampled subset of the RACE test set and Ceiling is the percentage of the unambiguous questions with a correct answer in a subset of the test set.

The overall performance is worse compared to the optimized versions of BERT (RoBERTa), trained with much larger corpus, but the perform substantially better than the majority baselines on both RACE-M and RACE-H. Furthermore, our proposed FNN-MRC method obtains significant improvement over two widely used neural models proposed in this paper, which leads to a 4.7%, 3.4% and 3.8% performance boost over the original BRET model on RACE-M, RACE-H and RACE-ALL, and 3.7%, 2.1% and 2.6% performance boost over bi-LSTM on RACE-M, RACE-H and RACE-ALL. Note we choose original BERT as our backbone, as it

Table 5
Parameters, FLOPs, latency and accuracy for different models on RACE.

Model	Parameters (M)	FLOPs (G)	Latency (ms)	Accuracy (%)
Bi-LSTM	1.8	3.2	63.9	47.7
Co-Match [39]	33.3	7.1	105.6	50.4
BERT-base [45]	110.0	174.0	293.1	65.0
FNN-MRC(BERT-base)	141.6	177.6	603.9	69.9
BERT-large [45]	335.1	618.5	963.1	72.0
FNN-MRC(BERT-large)	367.3	623.5	1219.2	75.8

Table 6
A case study example.

Passage	Katie went to the store...She looked around for the *flowers*. Katie then *looked* for the snacks. She wanted cookies not *chips*. She **found** some ***chocolate cookies***. Katie then looked for a *bow*....
Question	What snack did Katie **buy**?
Option	(A) Chips (B) Chocolate cookies (C) Flowers (D) Bows
Answer	B
Frame semantic	{Chips , Chocolate cookies} $\in$ Food {Flowers , Bows} $\notin$ Food **Found**, **Buy** and **looked** have relations, as their Frames are connected.

is widely used in many areas. Nevertheless, some other transformer models, which are optimized versions of BERT, can also be applied.

In addition, we also test the performance of our three different Frame representation methods (LUA, TLUA, FRA) on RACE. Table 4 shows their detailed results, where we observe the advantage of incorporating Frame semantics. Besides, the performance gained by adding Frame semantics is larger, signifying Frame semantics can incorporate more semantic knowledge into the model. FNN-MRC using different neural models (LSTM and BERT) outperforms themselves in all settings, 72.0% $\rightarrow$ 75.8% on BERT and 47.7% $\rightarrow$ 50.3% on LSTM. In fact, all the models using BERT outperforms the LSTM with significantly large margins, indicating pre-training model is very effective for learning semantics from unsupervised data.

Model complexity. It is commonly observed that the performance of deep neural networks is highly dependent on the model complexity, which is measured by the model size and computational consumption. Table 5 list the measured parameters, FLOPs and latency of different models on RACE. From Table 5, we can observe that: (1) A general trend is that the larger the model is, the higher accuracy it can achieve in a given task. For examples, BERT-base achieves 65.0% accuracy on RACE and the network contains about 110M parameters. BERT-large contains 335.1M parameters and significantly improves the accuracy to 72%. (2) Models with more parameters and more FLOPs will need more time to finish the computation and therefore will have higher latency. (3) Our FNN-MRC model is focusing on the model performance and not the model efficiency. It seems reasonable to reduce model sizes to achieve acceptable speed–accuracy balances and we leave it for future work.

6.5. Case study

For case study, Table 6 shows an example in MCTest, where our proposed model is able to answer it correctly. Note both *Chips*, *Chocolate cookies* belong to the ***Food*** Frame, while *Flowers* and *Bows* evoke two different Frames ***Plants*** and ***Accoutrements*** respectively. The target words **Found**, **Looked** and **Buy** in the given passage/question evoking different Frames ***Locationg***, ***Seeking*** and ***Commerce_buy*** — in FrameNet they are connected due to their semantic relations, as shown in Fig. 9, facilitating us to find answer B) Chocolate cookies.

6.6. Discussions

We observed that the proposed FNN-MRC model can utilize Frame semantic information to boost the MRC performance. From Tables 2 and 4, we can see that the Frame information makes a great contribution to the overall improvement, which confirms our hypothesis that Frame is useful for sentence understanding.

On the one hand, **Frame provides generalizations about lexical units at a useful level of abstraction.** As shown in Table 6, according to the word *snack* in question, FNN-MRC model focuses on *Chips* and *Chocolate cookies*, as all of them belong to the ***Food*** Frame.

On the other hand, **Frame relations provide a way to find semantic related sentences.** For example, for the question "*What snack did Katie **buy***" in Table 6, FNN-MRC model can help identify sentences *"Katie then **looked** for the snacks"* and *"She **found** some chocolate cookies"* with Frame relations information, which are very useful for answer selection.

Error Analysis. To better understand the performance of FNN-MRC model, we also conduct a manual error analysis and find two main types of samples that lead to the misclassification.

(1) Require world knowledge. These questions not only need knowledge contained in the given passage, but also need external world knowledge. For instance, Example 1 in Table 7, humans can easily know "*tomorrow*" is "*Saturday*" as "*today*" is "*Friday*". But it is very difficult for the model to find the correct answer as it does not have world knowledge.

(2) Require arithmetic operation. These questions require simple arithmetic operation over story elements to select the correct answer. For instance, Example 2 in Table 7, the question is "*How many friends did Little Bunny want to invite?*". It is challenging for the model since it must perform numerical addition, according to the sentences "*Bunny wanted to invite Rabbit, Bear, Duck and Goose*", "*He thought about Turtle*", "*He thought about Fox*" and "*all of your friends can have some*".

Limitations. Although FNN-MRC model performs well on some benchmark datasets, it still has some limitations.

(1) FNN-MRC model pays attention to Frame representation, but ignores critical Frame Elements representation, which is equally vital to MRC.

(2) Both *syntax* and *semantics* information of text are essential components for text understanding. FNN-MRC model focuses on modeling the *semantics* information while ignoring *syntax* structures.

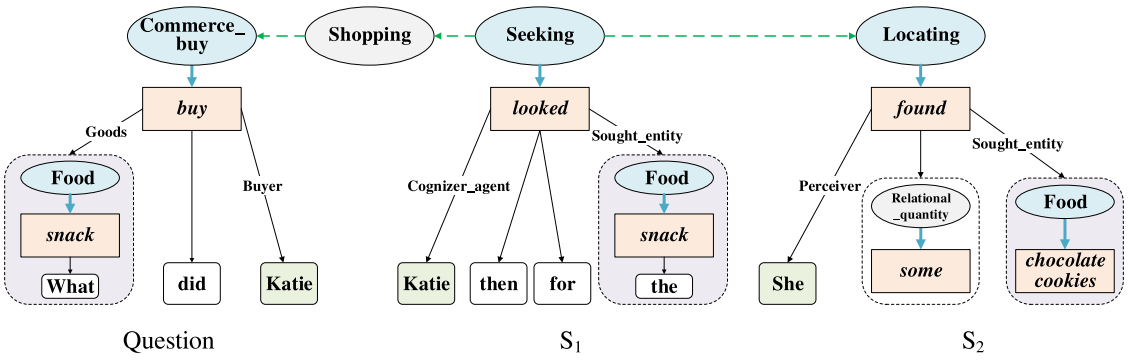

Fig. 9. An example from MCTest demonstrates the effectiveness of Frame semantic knowledge. Light orange rectangles and ellipses represent the target words and their corresponding Frames respectively. Dashed arrows (between Frames) are Frame-to-Frame relationships, and black arrows (between target words and words) represent Frame Elements.

Table 7
MCTest error cases. The correct answer is marked with ∗, and the answer guessed by the model is marked with †.

Example 1: Require world knowledge
Passage: It was a beautiful *Friday* morning in Los Angeles. Angela woke up and got dressed...She chose to make a salad for lunch *tomorrow* and Sunday. Question: What day was Angela making salad for lunch? (A) today ∗ (B) Saturday †(C) Sunday (D) Friday
Example 2: Require arithmetic operation
Passage: Tomorrow was Little Bunny's birthday. "We only have enough cake for five friends." His mother said. He wanted to invite *Rabbit*, *Bear*, *Duck* and *Goose*. Little Bunny could invite one more friend. He thought about *Turtle*. He thought about *Fox*... "I'll make a batch of cupcakes, and all of your friends can have some." Question: How many friends did Little Bunny want to invite? (A) 8 †(B) 5 ∗ (C)6 (D) 4

(3) FNN-MRC model performs poorly in cases requiring external knowledge or requiring deep comprehension, as shown in Table 7.

7. Conclusion and future work

In this paper, we proposed a novel Frame-based Neural Network for Machine Reading Comprehension (FNN-MRC). Specifically, we utilize both Lexical Units (LUs) and Frame-to-Frame (F-to-F) relations to built the Frame representation model, and propose a novel Frame-based sentence representation model to integrate multi-Frame semantic information in order to facilitate sentence modeling. Our extensive experimental results across four datasets demonstrate our proposed FNN-MRC works very well for the challenging machine reading comprehension tasks. Our error analysis suggests that incorporating world knowledge can yield further improvements on this task.

There are three interesting future research directions to extend our work. First, FNN-MRC model mainly focuses on Frame representation, which is relatively coarse-grained, and it is desirable to further design fine-grained semantic information representation method (e.g. Frame Elements representation). Second, our model can be improved by integrating external knowledge, such as structured information (like syntax) or word knowledge, to obtain richer and more comprehensive representation. Finally, it would be interesting to improve model efficiency with minimal quality degradation.

CRediT authorship contribution statement

Shaoru Guo: Conceptualization, Methodology, Software, Writing - original draft. **Yong Guan:** Software, investigation, Data curation. **Hongye Tan:** Resources, Validation. **Ru Li:** Supervision. **Xiaoli Li:** Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank the anonymous reviewers for their insightful comments. This work was sponsored by the National Natural Science Foundation of China (No. 61936012, No. 61772324).

References

[1] C.J. Fillmore, Frame semantics and the nature of language, Ann. New York Acad. Sci. 280 (1) (1976) 20–32, http://dx.doi.org/10.1111/j.1749-6632.1976.tb25467.x, URL https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.1976.tb25467.x, arXiv:https://nyaspubs.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-6632.1976.tb25467.x.

[2] C.F. Baker, C.J. Fillmore, J.B. Lowe, The berkeley framenet project, in: Proceedings of the 17th International Conference on Computational Linguistics, COLING '98, Association for Computational Linguistics, Stroudsburg, PA, USA, 1998, pp. 86–90, http://dx.doi.org/10.3115/980451.980860, https://doi.org/10.3115/980451.980860.

[3] D. Gildea, D. Jurafsky, Automatic labeling of semantic roles, Comput. Linguist. 28 (3) (2002) 245–288, http://dx.doi.org/10.1162/089120102760275983, https://doi.org/10.1162/089120102760275983.

[4] D. Das, D. Chen, A.F.T. Martins, N. Schneider, N.A. Smith, Frame-semantic parsing, Comput. Linguist. 40 (1) (2014) 9–56, http://dx.doi.org/10.1162/COLI_a_00163, https://doi.org/10.1162/COLI_a_00163.

[5] S. Liu, Y. Chen, S. He, K. Liu, J. Zhao, Leveraging framenet to improve automatic event detection, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 2134–2143, http://dx.doi.org/10.18653/v1/P16-1201, URL https://www.aclweb.org/anthology/P16-1201.

[6] A. Burchardt, M. Pennacchiotti, S. Thater, M. Pinkal, Assessing the impact of frame semantics on textual entailment, Nat. Lang. Eng, 15 (4) (2009) 527–550, http://dx.doi.org/10.1017/S1351324909990131.

[7] B. Ofoghi, J. Yearwood, L. Ma, The impact of frame semantic annotation levels, frame-alignment techniques, and fusion methods on factoid answer processing, J. Am. Soc. Inf. Sci. Technol. 60 (2) (2009) 247–263.
[8] N. Chambers, D. Jurafsky, A database of narrative schemas, in: N.C.C. Chair, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (Eds.), Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, ISBN: 2-9517408-6-7, 2010.
[9] X. Zhang, X. Sun, H. Wang, Duplicate question identification by integrating framenet with neural networks, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[10] G.A. Miller, Wordnet: A lexical database for english, Commun. ACM (ISSN: 0001-0782) 38 (11) (1995) 39–41, http://dx.doi.org/10.1145/219717.219748, https://doi.org/10.1145/219717.219748.
[11] M. Palmer, D. Gildea, P. Kingsbury, The proposition bank: An annotated corpus of semantic roles, Comput. Linguist. 31 (1) (2005) 71–106, http://dx.doi.org/10.1162/0891201053630264, URL https://www.aclweb.org/anthology/J05-1004.
[12] H. Wang, M. Bansal, K. Gimpel, D. McAllester, Machine comprehension with syntax, frames, and semantics, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Beijing, China, 2015, pp. 700–706, http://dx.doi.org/10.3115/v1/P15-2115, URL https://www.aclweb.org/anthology/P15-2115.
[13] K.M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, Teaching machines to read and comprehend, in: C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems 28, Curran Associates, Inc., 2015, pp. 1693–1701, URL http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend.pdf.
[14] D. Kapashi, P. Shah, Answering Reading Comprehension Using Memory Networks, Report for Stanford University Course cs224d, 2015.
[15] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, 2018, CoRR abs/1810.04805. arXiv:1810.04805.
[16] K.M. Hermann, P. Blunsom, Multilingual models for compositional distributed semantics, 2014, CoRR abs/1404.4641. arXiv:1404.4641.
[17] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Trans. Assoc. Comput. Linguist. 5 (2017) 135–146, http://dx.doi.org/10.1162/tacl_a_00051, https://doi.org/10.1162/tacl_a_00051.
[18] G. Glavas, R. Litschko, S. Ruder, I. Vulic, How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions, 2019, CoRR abs/1902.00508. arXiv:1902.00508.
[19] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2014, arXiv preprint arXiv:1409.0473.
[20] M. Richardson, C.J. Burges, E. Renshaw, MCTest: A challenge dataset for the open-domain machine comprehension of text, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle, Washington, USA, 2013, pp. 193–203, URL https://www.aclweb.org/anthology/D13-1020.
[21] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ questions for machine comprehension of text, 2016, arXiv preprint arXiv:1606.05250.
[22] A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, K. Suleman, NewsQA: A machine comprehension dataset, in: Proceedings of the 2nd Workshop on Representation Learning for NLP, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 191–200, http://dx.doi.org/10.18653/v1/W17-2623, URL https://www.aclweb.org/anthology/W17-2623.
[23] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735, https://doi.org/10.1162/neco.1997.9.8.1735.
[24] Y. Cui, T. Liu, Z. Chen, S. Wang, G. Hu, Consensus attention-based neural networks for Chinese reading comprehension, 2016, arXiv preprint arXiv:1607.02250.
[25] W.L. Taylor, "Cloze procedure": A new tool for measuring readability, Journalism Q. 30 (4) (1953) 415–433.
[26] M. Joshi, E. Choi, D.S. Weld, L. Zettlemoyer, Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension, 2017, arXiv preprint arXiv:1705.03551.
[27] G. Lai, Q. Xie, H. Liu, Y. Yang, E.H. Hovy, RACE: large-scale reading comprehension dataset from examinations, 2017, CoRR abs/1704.04683. arXiv:1704.04683.
[28] E. Smith, N. Greco, M. Bošnjak, A. Vlachos, A strong lexical matching method for the machine comprehension test, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 1693–1698, http://dx.doi.org/10.18653/v1/D15-1197, URL https://www.aclweb.org/anthology/D15-1197.
[29] K. Narasimhan, R. Barzilay, Machine comprehension with discourse relations in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 1253–1262.
[30] M. Sachan, E. Xing, Machine comprehension using rich semantic representations, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016, 486–492.
[31] M. Sachan, K. Dubey, E. Xing, M. Richardson, Learning answer-entailing structures for machine comprehension, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 239–249.
[32] J. Lu, J. Xuan, G. Zhang, X. Luo, Structural property-aware multilayer network embedding for latent factor analysis, Pattern Recognit. 76 (2018) 228–241.
[33] C. Li, Y. Wu, M. Lan, Inference on syntactic and semantic structures for machine comprehension, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[34] B. Dhingra, H. Liu, W.W. Cohen, R. Salakhutdinov, Gated-attention readers for text comprehension, 2016, CoRR abs/1606.01549. arXiv:1606.01549.
[35] D. Chen, J. Bolton, C.D. Manning, A thorough examination of the CNN/Daily mail reading comprehension task, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 2358–2367, http://dx.doi.org/10.18653/v1/P16-1223, URL https://www.aclweb.org/anthology/P16-1223.
[36] A. Trischler, Z. Ye, X. Yuan, J. He, P. Bachman, K. Suleman, A parallel-hierarchical model for machine comprehension on sparse data, 2016, CoRR abs/1603.08884. arXiv:1603.08884.
[37] Y. Xu, J. Liu, J. Gao, Y. Shen, X. Liu, Towards human-level machine reading comprehension: Reasoning and inference with multiple strategies, 2017, CoRR abs/1711.04964. arXiv:1711.04964.
[38] Y. Tay, L.A. Tuan, S.C. Hui, Multi-range reasoning for machine comprehension, 2018, CoRR abs/1803.09074. arXiv:1803.09074.
[39] S. Wang, M. Yu, J. Jiang, S. Chang, A co-matching model for multi-choice reading comprehension, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 746–751, http://dx.doi.org/10.18653/v1/P18-2118, URL https://www.aclweb.org/anthology/P18-2118.
[40] S. Wen, W. Liu, Y. Yang, P. Zhou, Z. Guo, Z. Yan, Y. Chen, T. Huang, Multilabel image classification via feature/label co-projection, IEEE Trans. Syst. Man Cybern. Syst. (2020) 1–10, http://dx.doi.org/10.1109/TSMC.2020.2967071.
[41] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 2227–2237.
[42] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R.R. Salakhutdinov, Q.V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, in: Advances in Neural Information Processing Systems, 2019, pp. 5754–5764.
[43] M. Joshi, D. Chen, Y. Liu, D.S. Weld, L. Zettlemoyer, O. Levy, Spanbert: Improving pre-training by representing and predicting spans, Trans. Assoc. Comput. Linguist. 8 (2020) 64–77.
[44] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019, CoRR abs/1907.11692. arXiv:1907.11692.
[45] S. Zhang, H. Zhao, Y. Wu, Z. Zhang, X. Zhou, X. Zhou, DCMN+: Dual co-matching network for multi-choice reading comprehension, 2019, arXiv preprint arXiv:1908.11511.
[46] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, B. Catanzaro, Megatron-lm: Training multi-billion parameter language models using model parallelism, 2019, arXiv–1909.
[47] X. Pan, K. Sun, D. Yu, J. Chen, H. Ji, C. Cardie, D. Yu, Improving question answering with external knowledge, 2019, arXiv preprint arXiv:1902.00993.
[48] B. Wang, S. Guo, K. Liu, S. He, J. Zhao, Employing external rich knowledge for machine comprehension, in: IJCAI, 2016, pp. 2925–2929.
[49] K. Sun, D. Yu, D. Yu, C. Cardie, Improving machine reading comprehension with general reading strategies, 2018, CoRR abs/1810.13441. arXiv:1810.13441.
[50] M. Kshirsagar, S. Thomson, N. Schneider, J.G. Carbonell, N.A. Smith, C. Dyer, Frame-semantic role labeling with heterogeneous annotations, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2015, pp. 218–224.

[51] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Empirical Methods in Natural Language Processing. EMNLP, 2014, pp. 1532–1543, URL http://www.aclweb.org/anthology/D14-1162.

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[53] B. Peng, Z. Lu, H. Li, K. Wong, Towards neural network-based reasoning, 2015, CoRR abs/1508.05508. arXiv:1508.05508.

[54] A. Radford, Improving language understanding by generative pre-training, 2018.

[55] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, 2019, CoRR abs/1907.11692. arXiv:1907.11692.