Relying on Discourse Analysis to Answer Complex Questions by Neural Machine Reading Comprehension

Boris Galitsky
Oracle Inc. /
Redwood Shores, CA, USA
boris.galitsky@oracle.com

Dmitry Ilvovsky
National Research University
Higher School of Economics /
Moscow, Russia
dilvovsky@hse.ru

Elizaveta Goncharova
National Research University
Higher School of Economics /
Moscow, Russia
egoncharova@hse.ru

Abstract

Machine reading comprehension (MRC) is one of the most challenging tasks in natural language processing domain. Recent state-of-the-art results for MRC have been achieved with the pre-trained language models, such as BERT and its modifications. Despite the high performance of these models, they still suffer from the inability to retrieve correct answers from the detailed and lengthy passages. In this work, we introduce a novel scheme for incorporating the discourse structure of the text into a self-attention network, and, thus, enrich the embedding obtained from the standard BERT encoder with the additional linguistic knowledge. We also investigate the influence of different types of linguistic information on the model's ability to answer complex questions that require deep understanding of the whole text. Experiments performed on the SQuAD benchmark and more complex question answering datasets have shown that linguistic enhancing boosts the performance of the standard BERT model significantly.

1 Introduction

Machine reading comprehension (MRC) reflects the ability to read and understand an unstructured text and answer questions regarding it. Aiming to find the relevant answer to a question in the form of a text span, the MRC models should demonstrate deep understanding of the language and text organization.

Transformer models that achieve state-of-the-art results on multiple natural language processing (NLP) tasks have been successfully applied to the MRC. However, while the ideal MRC model should read most words superficially and pay attention only to the essential ones (Wang et al., 2017), the attention mechanism in the standard transformers attends to all words without explicit constraint which results in inaccurate concentration on some less important text spans. Lately, the researchers have actively examined the ability of the deep learning (DL) models to understand language and build accurate linguistic-enhanced internal representation.

Recent works have revealed that traditional DL models that ignore additional linguistic knowledge, such as syntax or semantic, achieves lower accuracy on such complex tasks as natural language understanding (NLU) or MRC (Roth and Lapata, 2016; Marcheggiani and Titov, 2017; He et al., 2018). It has been shown that incorporating explicit syntactic (Hu et al., 2019) and semantic (Zhang et al., 2020a) relations into the attention mechanism leads to better linguistically motivated word representations beneficial for the MRC task.

Moreover, providing exact, concise answers frequently requires not just syntactic/meaning similarity but an overall structure of thoughts expressed by an author (Galitsky et al., 2013), i.e., some claims introduced by an author and logical connections existing among them. This information is encoded by discourse structure of a text that, as long as syntax and semantic, is believed to provide valuable information that could help the model to capture all the hidden dependencies existing in the text and to pay attention only to the relevant words while answering the corresponding question.

In this paper, we explore if and how discourse-level features (discourse relations connecting the text spans), fed to a neural MRC model on top of syntactic and semantic features or independently, can help to answer complex, long, multi-sentence questions. We intend to develop a neural method

Proceedings of Recent Advances in Natural Language Processing, pages 444–453
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_051

that selects relevant words by only considering the related subset of words, w.r.t. syntactic, semantic, and discourse-level importance. To provide feature encoding we use a self-attention network (SAN) enriched with the discourse features (such as *explanation*, *condition*, etc.) retrieved from a text and combine it with the classical transformer encoder to build linguistically-enhanced text representation.

Overall, the contribution of this paper is threefold: first, we introduce a novel discourse-aware transformer-based model to construct the enriched internal representation of the text. Second, we develop an ensemble MRC model that combines syntax, semantic, and discourse MRC components. Third, we conduct experiments on various question-answering (QA) datasets to assess the ability of the linguistically enriched model to answer complex questions and estimate the influence of each source of linguistic information.

2 Related Work and Background

2.1 Machine Reading Comprehension

Span-based MRC, which is the main focus of this work, is quite a challenging task, as we expect the model not only to identify the relevant document that contains a possible answer but to retrieve the exact text fragment that answers the question. There has been a lot of studies on solving this task with attentive models (Kadlec et al., 2016; Yuan et al., 2018; Guo et al., 2019).

Recently, the pre-trained contextual language models (LMs) such as ELMO (Peters et al., 2018), BERT (Devlin et al., 2019), or a series of GPTs (Radfort et al., 2018) have shown state-of-the-art results on the number of NLU benchmarks which has attracted the researchers' interest toward utilizing these models for MRC. Despite the increasing popularity of these LMs, several studies have revealed that textual representation provided by them relies purely on the context of each word and, generally, neither the syntactical nor semantic organization of the text is considered. As this information is crucial for MRC, the novel techniques to incorporate syntactic and semantic knowledge into the pre-trained LMs have been the main focus of the latest works.

2.1.1 Syntactic-aware Models

Recent attempts to turn neural network algorithms into more structure-aware ones have discovered the incorporation of external memories in the context of recurrent neural networks. The idea is to use multiple memory slots outside the recurrence to piecewise store representations of the input. Read and write operations for each slot can be modeled as an attention mechanism with a recurrent controller. Cheng et al. (2016), for example, leverage memory and attention to empowering a recurrent network with stronger memorization capability and more importantly the ability to discover relations among tokens. This is realized by inserting a memory network module in the update of a recurrent network together with attention for memory addressing. The attention acts as a weak inductive module discovering relations between input tokens and is trained without direct supervision. The experiments performed on NLI datasets showed that the superiority of the modified model over the vanilla LSTMs.

In more recent work (Zhang et al., 2020b), the authors benefit from the performance of the BERT model on span-based MRC tasks and sponsor it with the syntax-guided SAN. They design an informative method that can selectively pick out important words by only considering the related subset of syntactically important context inside each input sentence explicitly. With the guidance of syntactic structure clues, the syntax-guided method could give more accurate attentive signals and reduce the impact of the noise brought about by lengthy sentences. The authors extend the self-attention mechanism with syntax-guided constraint, to capture syntax-related parts with each concerned word. Specifically, they adopt a pre-trained dependency syntactic parse tree structure to produce the related nodes for each word in a sentence, namely *syntactic dependency of interest*, by regarding each word as a child node, and the syntactic dependency of interest consists of all its ancestor nodes and itself in the dependency parsing tree. The syntax encapsulating into the model should provide a better understanding of the long or unanswerable questions, which is a big obstacle for the existing MRC models.

2.1.2 Semantic-aware Models

Frequently, DL models suffer from insufficient contextual semantic representation and learning. So, the way of constructing semantic-aware LMs has also attracted wide attention in research.

To provide contextual semantic representation to the DL models, Strubell et al. (2020) propose linguistically-informed self-attention (LISA), which is used for the semantic role labeling (SRL) task. The model proposed by the authors is end-to-end, and it is trained to predict part of speech

tags, provide parsing, attend to syntactic parse dependents, and, finally, assign semantic role labels to the model. This architecture has been applied to enlarge the contextual representation provided by BERT with the additional semantic information.

In (Zhang et al., 2020a), the authors propose to use SRL task to integrate the text representation provided by BERT with the contextual explicit semantic embedding, the introduced model is called *Sem-BERT*. Sem-BERT is intended to handle multiple sequence inputs, the words in the input sequence are passed to semantic role labeling to obtain multiple predicate-derived structures to form a semantic embedding. In parallel, the input sequence is segmented to subwords (if any) by BERT word-piece tokenizer, then the subword representation is transformed back to word-level via a convolutional layer to obtain the contextual word representations. Finally, the word representations H and semantic embedding H'_{sem} are concatenated to form the joint representation.

Despite there is a number of works encapsulating the syntactic and semantic information about the text into the DL models, there is still a lack of research that considers discourse organization, which also introduces relevant linguistic knowledge essential for MRC, and other downstream challenges. In this work, we propose a way to encode the discourse structure of the text by neural network and enrich the text embeddings constructed by BERT with this information. Then, we aim to assess the influence of discourse, semantic, and syntactic features on the MRC task.

2.2 Discourse Structure

In this section, we introduce the definition of discourse structure that we propose to integrate into the MRC model. Any coherent text is structured so that we can derive and interpret the information. This structure shows how discourse units (text spans such as sentences or clauses) are connected and relate to each other. Discourse analysis reveals this structure and describes the relations that hold between text units in the document. Several theories have been proposed in the past to describe the discourse structure, among which the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) is one of the most popular. RST divides a text into minimal atomic units, called Elementary Discourse Units (EDUs). It then forms a tree representation of discourse called a Discourse Tree (DT) using rhetorical relations (Elaboration, Explanation, etc.) as edges, and EDUs as leaves. EDUs linked by a rhetorical relation are also distinguished based on their relative importance in conveying the author's message: the nucleus is the central part, whereas the satellite is the peripheral part. Nucleus units consist of basic information and satellite units contain additional information about the nucleus.

An exploration of coherence relations in frameworks such as RST has experienced a revival in the decade in English and a few other languages (Matthiessen and Teruya, 2015; Maziero et al., 2015; Zeldes, 2016) which has led to a grown number of applications of discourse analysis. For example, discourse parsers are used in argumentation mining in online discussions, summarization, QA systems, and machine translation (Benamara et al., 2017; Durrett et al., 2016; Peldszus and Stede, 2016; Chakrabarty et al., 2020). We claim that incorporating this additional discourse information provided by state-of-the-art parsers could be beneficial for DL models performing MRC. We are motivated to improve the self-attention layer appended to the top of the transformer encoder to enrich the contextualized word representation with information from its neighbors and the relations from the dependency parse trees.

3 MRC System Extended with Discourse Relations

In this paper, we present the novel discourse-aware attentive model designed to perform the MRC task. Our approach is inspired by syntax-guided BERT (Zhang et al., 2020b), while instead of encapsulating the syntactic dependencies among the words, we pre-process the discourse parse tree and observe the EDUs as long as the specific discourse relations connecting them.

We introduce the architecture of the transformer-based encoder empowered with the discourse knowledge about the input text in Section 3.1. As we aim to assess the influence of all three types of linguistic information, in Section 3.2, we present the final MRC system designed as the ensemble of state-of-the-art syntactic, semantic, and the proposed discourse-aware attention components.

3.1 Discourse-aware Model

In this section, we describe a method for incorporating discourse relations into the transformer-based model explicitly. As well as the syntactic

dependency parse tree, the discourse structure can be represented as a hierarchically organized tree, where the leaves are the text spans and the edges denote the type of relations connecting them. Thus, we propose to modify the SAN appended to the top of vanilla transformer-based encoder to make it able to process the discourse text organization, and, thus, to utilize this additional linguistic feature for MRC.

3.1.1 Discourse-aware Self-attention Layer

Our discourse-aware language model is trained to provide the vector representation of the text enriched by the discourse relations connecting text units. To obtain this representation we use the standard transformer encoder to calculate contextual representation of the text, then the obtained vector is passed through the discourse-aware SAN, which is designed to encapsulate the discourse structure into the embedding of the sequence. Finally, the discourse-aware representation is aggregated with the output of the pure transformer, this final embedding goes through the task-specific layer to perform the MRC task. The overall model architecture is presented in Fig. 1.

Generally, the main difference between the discourse-aware language model and the traditional transformer-based model is as follows. In traditional transformers, the word attends to both sides of the context, while in the discourse-aware model we would like each word to attend to its discourse-dependent ancestors. This forces a multi-head attention mechanism to analyze the dependency among tokens w.r.t. the rhetoric relations connecting them. As we have already mentioned, the discourse structure of the text is represented by the DT. In this section, we will present the approach for incorporating this DT into the SAN.

To provide the discourse structure of the text we use a state-of-the-art discourse parser (Joty et al., 2013) which constructs a hierarchically organized dependency tree for the input text. The text annotated with discourse relations will be transmitted to the attention network. Whereas the SAN cannot encapsulate the whole discourse tree, we need to detect the essential dependencies existing among words that should be included in the model. Each discourse unit (sequence of words) corresponding to some leaf in DT is connected to its ancestor non-terminal node labeled by the rhetoric relation referred to this EDU. For example, in the passage and its DT shown in Fig. 3, the words do not attend to all of the left and right neighbors in the context, on the contrary, the words *finds* and *clinicians* are connected to their ancestor labeled by *Attributions*, while the *pneumonia* attends to its ancestor *Cause* which also depends on *Attributions*. This sequence of connections fully reflects the discourse organization of the text.

Formally, given input token sequence $S = \{s_1, s_2, ..., s_n\}$ of length n, we first pass it through the discourse parser to split into the EDUs and generate the discourse dependencies existing between them. The input sequence after parsing is enriched with the discourse relations $\overline{S}_{rel} = \{rel_1, edu_1, rel_2, edu_2, ..., rel_m, edu_m\}$ of length m where $edu_i = \{s_k, s_{k+1}, .., s_{k+K}\}$, and K is the number of tokens assigned to the ith EDU. We should notice that edu_i could be an empty set if rel_i connects two non-terminals nodes corresponding to the sub-trees in the DT (see *contrast→elaboration* relations in Fig. 3). Then, we should retrieve the ancestor nodes for each of the word s_i and the rhetoric relation rel_i. To provide this we traverse the discourse dependency tree, and the ancestor node set P_i is derived for each s_i and rel_i. Finally, in an analogy with syntax-guided SAN a discourse dependency of interest mask M is obtained. M is $(n+m) \times (n+m)$ matrix, where the elements in each row denote the dependency mask of all tokens to the row-index token. $M[i,j] = 1$ means that token s_i is the ancestor node of token s_j.

$$M[i,j] = \begin{cases} 1 & \text{if } j \in P_i \text{ or } j = i \\ 0 & \text{otherwise.} \end{cases}$$

To obtain the discourse-aware representation of the text we project the last layer output H of size L calculated by the original transformer encoder into the distinct key, value, and query representations of dimensions $< L \times d_k, L \times d_q, L \times d_v >$, respectively, denoted $< K'_i, Q'_i, V'_i >$ for each headword i. Then a dot product is computed to score key-query pairs with the dependency of interest mask to obtain attention weights of dimension $L \times L$, denoted A'_i:

$$A'_i = Softmax(\frac{M\left(Q'_i K'^T_i\right)}{\sqrt{d_k}})$$

The attention weight A'_i is multiplied by V'_i to obtain the discourse-aware token representations: $W'_i = A'_i V'_i$. W'_i for all heads are concatenated and passed through a feed-forward layer. After passing through another feed-forward layer, a layer

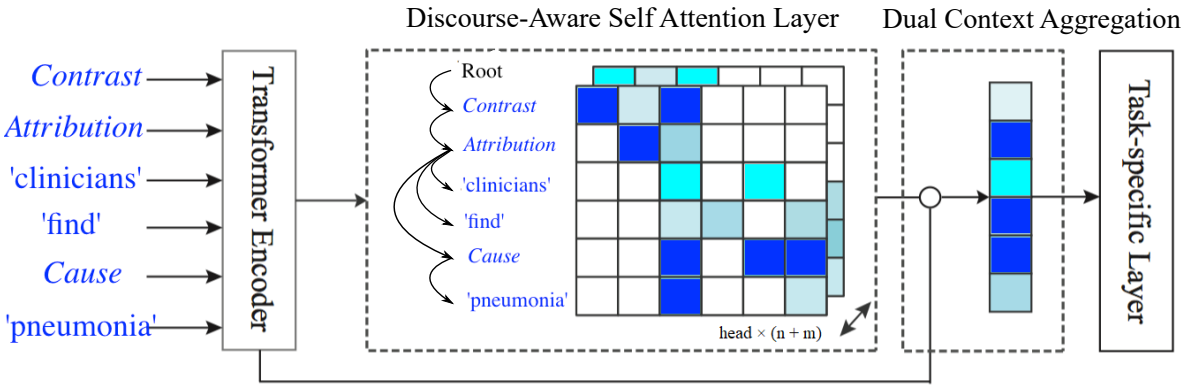

Figure 1: Architecture of the discourse-aware model.

normalization is applied to the sum of output and initial representation to obtain the final $H'_{disc} = \{h'_0, h'_1, ..., h'_n\}$.

Finally, we summarize the two text representations, where the former is obtained from the standard transformer encoder H, and the latter is the discourse-aware text representation H'_{disc}, finally $\overline{H}_{disc} = H + H'_{disc}$.

3.1.2 Answer Detection

Having identified the model for calculating discourse-aware text representation, we could proceed with the MRC task. MRC is the ability to answer the question based on the input paragraph of the text. As we have already mentioned, in this work, we consider a so-called *span-based* MRC, where the answer should be found as the span of the input passage referring to the question. Formally, we can define the span-based MRC by a triple $< P, Q, A >$, where P is the text paragraph which is the basis for the question Q, and A is the correct answer to the question.

The input data which is fed to the transformer encoder is performed as *[CLS] P [SEP] Q [SEP]*, where the [CLS] and [SEP] are the special tokens utilized in the BERT model.

We use BERT model as the transformer encoder, so, the [CLS] token representation calculated by the BERT encoder for the input sequence is used as the contextualized representation H of the whole text passage and question. Finally, H goes through the linguistically enriched SAN in order to obtain H'_{ling} and $\overline{H}_{ling}$, where $ling \in [synt, sem, disc]$ that refers to the syntax, semantic, and discourse-aware SAN, respectively. $\overline{H}_{ling}$ is fed to a linear layer to obtain the probability distribution over the start and end positions of the answer in the text through a softmax layer.

In the work, we propose to analyze the influence of various linguistic characteristics on the MRC. So, for the experiments, we will use both the standalone $\overline{H}_{ling}$, and their combination, calculated as the sum of the individual $\overline{H}_{ling}$.

3.2 MRC Pipeline

Fig. 2 demonstrates the architecture of the whole pipeline that we introduce to perform the MRC task. All in all the main components of the model are as follows:

1. Linguistic data preparation, which extracts, organizes, and aligns linguistic features at various levels of knowledge abstraction. There the system parses the input text passage to obtain relevant linguistically enriched structures that will be further utilized in neural model performance. Discourse after-parser is responsible for enrichment of the input sequence with the discourse relations revealed form the DT, $\overline{S}_{rel}$.

2. Deep learning component actually performing MRC. This block provides encoding of the input passage and related questions using the classical transformer encoder as long as the additional linguistic feature extraction. The output of the context aggregation block is the representation $\overline{H}_{ling}$, which is the sum of context-based text embedding and the embedding provided by linguistically-guided SAN.

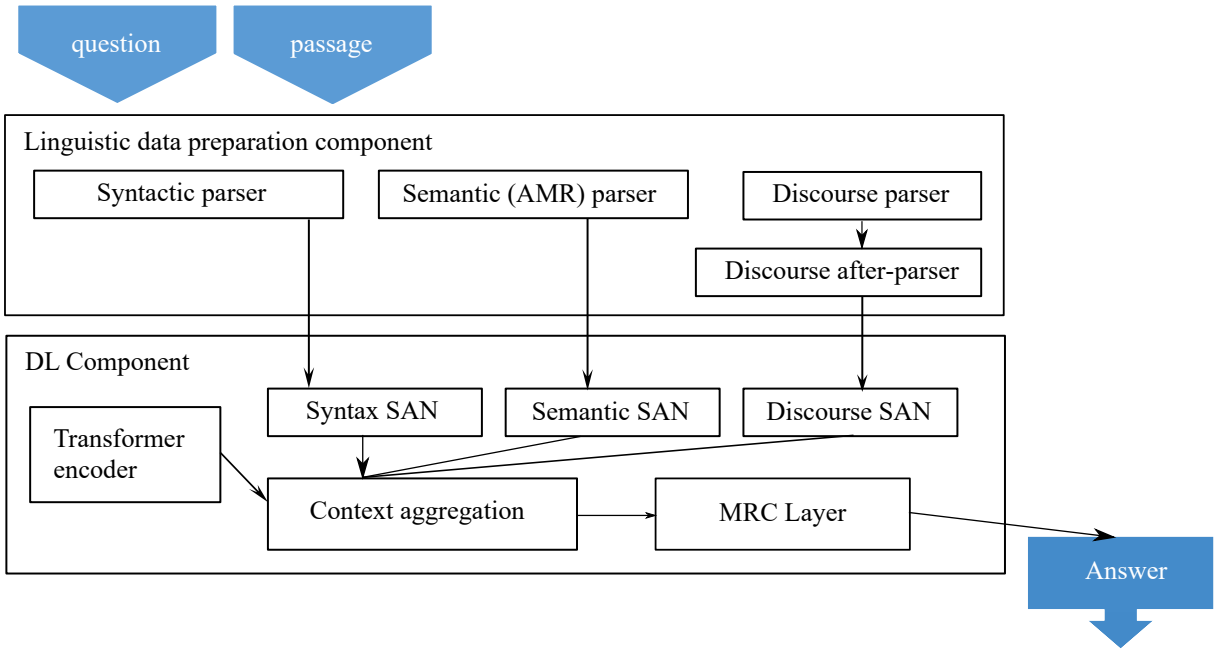

Figure 2: Architecture of linguistically-enabled MRC system.

Finally, this $\overline{H}_{ling}$ is used to perform the MRC task.

The discourse SAN block is the one introduced in Section 3.1.1. To provide the syntactic- and semantic-aware models we use the state-of-the-art models described in Sections 2.1.1 and 2.1.2. Specifically, we implement syntax-guided SAN by (Zhang et al., 2020b) and Sem-Bert by (Zhang et al., 2020a). These models are able to encapsulate the corresponding linguistic features into the transformer-based models that help to achieve an accuracy gain for the tasks related to MRC.

4 Experiments

In this work, we rely on four QA datasets with long, complex, multi-hop questions to observe if/how syntactic, semantic, and, mainly, discourse-level features help to provide the correct answers. As the baseline, we use fine-tuned BERT model. Besides, we compare the performance of our system with the current state-of-the-art results published or obtained from the leaderboard for the corresponding dataset.

4.1 Datasets and Setup

The experimental evaluation has been performed on several extracting reading comprehension English datasets. First, we verified the model on the well-known SQuAD datasets (Rajpurkar et al., 2016, 2018). then we evaluated how the introduced MRC model can cope with the more complex questions that require language comprehension and understanding of the full text rather than just a small paragraph. As the example of complex questions datasets, we consider NewsQA (Trischler et al., 2017), QA in Context (QuAC) dataset (Choi et al., 2020), and multi-sentence questions (MSQ) (Burchell et al., 2020).

Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowd-workers on a set of Wikipedia articles, where the answer to every question is a segment of text from the corresponding reading passage. SQuAD contains more than 100,000 question-answer pairs on 500 articles, which is significantly larger than previous reading comprehension datasets. We use two versions of this corpus: SQuAD 1.1 and SQuAD 2.0, where the latter also includes unanswerable questions so that we can test the ability of the model to detect the questions that cannot be answered based on the provided paragraph. F1 score that measures the weighted average of the word-level precision and recall rate is used to evaluate the performance of the models.

NewsQA dataset consists of 100K QA pairs written by humans for CNN news articles. Answers are typically the multiword spans of the source text, as in the SQuAD there are unanswerable questions

presented. The main challenge of this dataset is that a significant proportion of questions cannot be solved without reasoning, i.e. understanding conceptual overlap or identifying the synonyms.

MSQ dataset uses the Stack Exchange Data Explorer, an open-source tool for running arbitrary queries against public data from the Stack Exchange network. The authors of this corpora chose 93 sites within the network and queried each site for entries with at least two question marks in the body of the question. Also, the authors filtered too short (under 5 characters) and too long (over 300 characters), and badly formed questions. After cleaning and processing, 162,745 questions from 93 topics were extracted. This dataset includes the questions that consist of several sequential questions, and in order to answer them right, they should be considered as the one. We are not aware of any works that has attempted to improve QA performance on MSQs so far.

QuAC dataset has 100K QA pairs created by two crowd workers who are asking and answering questions about a hidden Wikipedia text. This dataset is aimed at enabling the MRC model to answer the latest question by comprehending not only the given context passage but all the dialogue that has been seen so far.

4.2 Results

To assess the influence of different linguistic features on the model performance we divided our experiments into two parts. Firstly, we provide the results on SQuAD datasets, then we present the evaluation on the more complex (w.r.t. the questions' design) NewsQA, QuAC, and MSQ datasets. In all experiments, we calculate F1 score as the weighted average between precision and recall. The results achieved by the introduced MRC model are presented in the bottom block of the table. We also show the results of the state-of-the-art models presented in the literature or public leaderboards (* symbol is used to refer to the unpublished works) for the available datasets (upper block). The results achieved by the MRC models relying on discourse information are in bold.

SQuAD. Our performance on both SQuAD 1.1 and 2.0 test data is shown in Table 1. The default MRC (baseline) employs neither syntactic nor semantic information, this is a typical fine-tuned cased BERT used as the encoder for the question and the passage. As we move towards syntactic, semantic, and discourse levels the average performance gain is 2.2, 3.4, and 3% respectively. The improvement of the integrated system is 5.4%. Despite the fact that the introduced model could not outperform the best both single (such as ALBERT) and ensemble (FPNet) models, we can observe that it boosts the default linguistic-free baseline essentially.

Dataset/settings	v1.1 test	v2.0 test
	F1	**F1**
SQuAD leaderboard		
FPNet*	-	93.18
Retro-Reader (Zhang et al., 2020c)	-	92.98
ALBERT (Lan et al., 2020)	-	92.20
LUKE*	95.4	-
Baseline	88.61	83.98
Syntax MRC	89.90	87.13
Semantic MRC	90.60	88.76
Discourse MRC	**90.08**	**88.60**
Syntax w. semantic w. discourse MRC	**93.14**	**90.20**

Table 1: F1 scores (%) on SQuAD 1.1 (v1.1) and SQuAD 2.0 (v2.0) datasets.

Dataset/settings	NewsQA	QuAC	MSQ
	F1	**F1**	**F1**
literature + QuAC leaderboard			
SpanBERT (Joshi et al., 2020)	73.6	-	-
DecaProp (Tay et al., 2018)	66.3	-	-
RoR*	-	74.9	-
FlowQA (Huang et al., 2019)	-	64.1	-
Baseline	66.48	65.69	60.66
Syntax MRC	70.95	71.09	66.79
Semantic MRC	71.84	70.15	66.55
Discourse MRC	**72.13**	**72.40**	**67.80**
Syntax w. semantic w. discourse MRC	**75.05**	**74.88**	**71.65**

Table 2: F1 scores (%) on complex questions datasets. The performance of other MRC models on MSQ dataset has not been published yet.

Complex datasets. Table 2 shows the result on NewsQA, QuAC, and MSQ. As we proceed to-

wards evaluation in the datasets of more complex questions, the performance drops up to 20%. Analogously to Table 1, the default MRC employs none of the additional linguistic information. Whereas the absolute performance value is lower than in Table 1, the performance boost due to linguistic information is higher. The average contributions of syntactic, semantic, and discourse levels are 5.3, 5.2, and 6.5% respectively. One can observe that contribution of discourse-level features is the highest in this evaluation domain of longer, multi-sentence questions (MSQ). The improvement of the integrated system is almost 11% for MSQ, and 9.5% on average. Hence, the more long and complex the questions are, the higher the impact of linguistic information, especially discourse-level. We should also mention that the introduced ensemble model outperforms both the stand-alone fine-tuned BERT and current state-of-the-art models for NewsQA and achieves comparable results on QuAC.

4.3 Case Study

Finally, let us consider a case study, where the linguistic-free BERT model provides the wrong result answering the question, while the introduced discourse-aware MRC model can answer the question correctly. Bellow, there are an input passage and the question regarding it.

P: ***Viruses, bacteria, and fungi*** *can all cause pneumonia. In the United States, common causes of viral pneumonia are influenza and respiratory syncytial virus. A common cause of bacterial pneumonia is Streptococcus pneumonia. However, clinicians are not always able to find out which germ caused someone to get sick with pneumonia.*

Q: *Who experience difficulties finding causes for pneumonia?*

The answer found by ELMO is *Viruses, bacteria, and fungi*, which, indeed, is not correct. The correct answer is *clinicians*. MRC fails miserably here associating *virus, bacteria, and fungi* with *Who*. Also, MRC failed to match the question with the sentence "However, clinicians are not always able to find out which germ caused someone to get sick with pneumonia.". The introduced discourse-aware model answers the question as "Clinicians are not always able to find out.". Let us consider the discourse structure for the passage and the question to understand the influence of discourse knowledge while dealing with this example.

The DT for this passage is shown in Fig. 3. In accordance with the constructed DT, we have a mapping between: *Q*: attribution → *P*: attribution, *Q*: cause → *P*: cause, *Q*: "causes" → *P*: "caused". This information allows the model to attend each word to the relevant text spans in the input passage and, thus, to find the correct answer to the question.

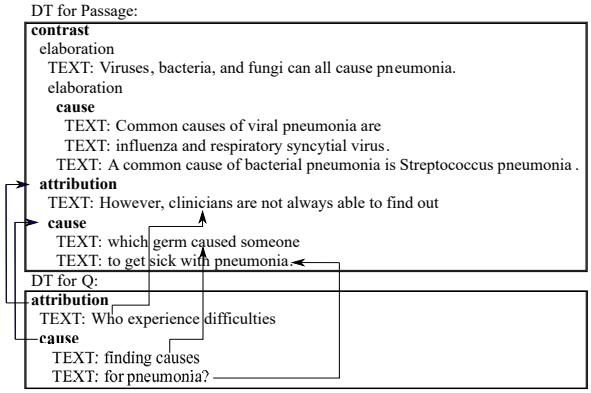

Figure 3: The discourse tree (DT) for text to chose an answer from (on the top) and for the question (on the bottom) with the mappings between corresponding nodes.

5 Conclusion

In this paper, we analyzed various linguistically enriched deep neural models and assessed the influence of semantic, syntax, and discourse on their performance on MRC tasks. While, modern systems are usually linguistic-free or rely on some independent linguistic characteristic, such as syntax or semantic individually, we claim that their combination could provide even higher accuracy gain. We also introduce the approach to incorporate discourse structure into the transformer-based model, which has been proven to be necessary for answering complex multi-sentence questions.

We have shown that the combination of three additional features encoded into a neural MRC is able to answer lengthy and complex questions better than the linguistic-free models, even the ones fine-tuned on the observed datasets. The introduced discourse-aware MRC model outperformed standalone syntax-guided (Zhang et al., 2020b) and semantic-enhanced models (Zhang et al., 2020a) for all the observed datasets. Although our MRC system did not achieve state-of-the-art results on some of the evaluation datasets (e.g., on the SQuAD), it demonstrated the superiority of integrated syntax/semantic/discourse subsystems in multiple diverse QA domains with complex questions.

References

Farah Benamara, Maite Taboada, and Yannick Mathieu. 2017. Evaluative language beyond bags of words: Linguistic insights and computational applications. *Computational Linguistics*, 43.

Laurie Burchell, Jie Chi, Tom Hosking, Nina Markl, and Bonnie Webber. 2020. Querent intent in multi-sentence questions. arXiv:2010.08980.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathleen McKeown, and Alyssa Hwang. 2020. AmperSand: Argument mining for persuasive online discussions. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*.

Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. In *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen Tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2020. QUAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, volume 4.

Boris A. Galitsky, Sergei O. Kuznetsov, and Daniel Usikov. 2013. Parse thicket representation for multi-sentence search. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7735 LNCS.

Jiabao Guo, Gang Liu, and Caiquan Xiong. 2019. Multiple attention networks with temporal convolution for machine reading comprehension. In *ICEIEC 2019 - Proceedings of 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication*.

Luheng He, Kenton Lee, Omer Levy, and Luke Zettlemoyer. 2018. Jointly predicting predicates and arguments in neural semantic role labeling. In *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, volume 2.

Minghao Hu, Furu Wei, Yuxing Peng, Zhen Huang, Nan Yang, and Dongsheng Li. 2019. Read + verify: Machine reading comprehension with unanswerable questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33(01).

Hsin Yuan Huang, Wen Tau Yih, and Eunsol Choi. 2019. FlowQA: Grasping flow in history for conversational machine comprehension. In *7th International Conference on Learning Representations, ICLR 2019*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Yashar Mehdad. 2013. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, volume 1.

Rudolf Kadlec, Martin Schmid, Ondřej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, volume 2.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. arXiv:1909.11942.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*.

Christian M.I.M. Matthiessen and Kazuhiro Teruya. 2015. Grammatical realizations of rhetorical relations in different registers. *Word*, 61.

Erick G. Maziero, Graeme Hirst, and Thiago A.S. Pardo. 2015. Semi-supervised never-ending learning in rhetorical relation identification. In *International Conference Recent Advances in Natural Language Processing, RANLP*, volume 2015-January.

Andreas Peldszus and Manfred Stede. 2016. Rhetorical structure and argumentation structure in monologue text. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1.

Alec Radfort, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. (OpenAI transformer): Improving language understanding by generative pre-training. *OpenAI*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, volume 2.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*.

Michael Roth and Mirella Lapata. 2016. Neural semantic role labeling with dependency path embeddings. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, volume 2.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2020. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*.

Yi Tay, Luu Anh Tuan, Siu Cheung Hui, and Jian Su. 2018. Densely connected attention propagation for reading comprehension. In *Advances in Neural Information Processing Systems*, volume 2018-December.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset.

Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2017. Learning sentence representation with guidance of human attention. In *IJCAI International Joint Conference on Artificial Intelligence*.

Hang Yuan, Jin Wang, and Xuejie Zhang. 2018. YNU-HPCC at Semeval-2018 task 11: Using an attention-based CNN-LSTM for machine comprehension using commonsense knowledge.

Amir Zeldes. 2016. rstWeb - a browser-based annotation interface for rhetorical structure theory and discourse relations.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020a. Semantics-aware BERT for language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34.

Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. 2020b. SG-Net: Syntax-guided machine reading comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2020c. Retrospective reader for machine reading comprehension. arXiv:2001.09694.