科学技術論文に対する談話依存構造のアノテーション・ガイドライン (簡易版)

西田 典起
理化学研究所 革新知能統合研究センター 知識獲得チーム

noriki.nishida@riken.jp

1 はじめに

本ガイドラインでは、本コーパス構築の目的と、アノテーション仕様、談話依存構造の詳細について説明する。なお、本稿の例示等で用いているアノテーションツールについては以下のサイトで公開している: https://norikinishida.github.io/tools/discdep/.

2 本コーパスの目的

談話依存構造 (Discourse Dependency Structure) は、文章中の節 (clause) や文の間の関係 (背景、手段、目的、例示など) をグラフ構造で表したものである。談話依存構造の例を図 1に示す。談話構造は、文書要約や商品レビューの極性分類、質問応答などの自然言語処理タスクで有用であることが知られている。

科学技術論文に対して談話依存構造をアノテーションした既存のコーパスとして、SciDTB [3] がある。SciDTB は、自然言語処理分野の論文のアブストラクト 798 件に対して、人手で談話依存構造のアノテーションを行っている。

論文の書き方や論旨の展開方法、用いられる語彙等は分野によって大きく異なる。実際に、私たちは最近、SciDTB を用いて学習した談話構造解析モデルを医学生物学分野の論文アブストラクトに適用すると、解析精度が著しく低下することを発見した。これは、機械学習では「ドメイン適応」の問題として知られている。

本コーパス構築の目的は、談話依存構造解析におけるドメイン (分野) 適応の研究のための評価用データセットを獲得することである。医学生物学分野を対象にし、特に COVID-19 および関連ウイルス (e.g., SARS) についての論文のアブストラクト約 200 件に対して、談話依存構造をアノテーションする。将来的には、本コーパスを用いて、談話依存構造に基づく新たな医療知識の発見方法の開発に繋げたい。

3 アノテーション概要

一般の談話構造のアノテーションは、

1. 談話単位分割
2. 談話構造解析

の 2 つのステップによって行われる。

最初の談話単位分割では、その名の通り、対象のテキストを談話単位 (Elementary Discourse Unit; EDU) と呼ばれる最小のテキスト領域に分割する。次の談話構造解析では、各 EDU 間の依存性の有無を同定し、そして依存性があるならばリンクを張り、どのような談話関係によって結合されるのかをラベル付けする。

今回、医学生物学論文のアブストラクト 200 件に対するEDU 分割については既に済んでおり、それらに対して談話依存構造をアノテーションする。EDU 分割済みデータは、EDU が 1 行ずつ書かれたテキストファイル (*.txt) として保存されており、それをアノテーションツールにアップロードすることによって談話依存構造のアノテーションを開始することができる。アノテーション結果は、アノテーションツールによって出力される JSON データとして保存する。

談話依存構造については次章で詳しく説明する。

4 談話依存構造

本節では、談話依存構造において重要な要素である

1. 談話単位 (EDU)
2. 依存構造
3. 談話関係

のそれぞれについて簡単に説明する。

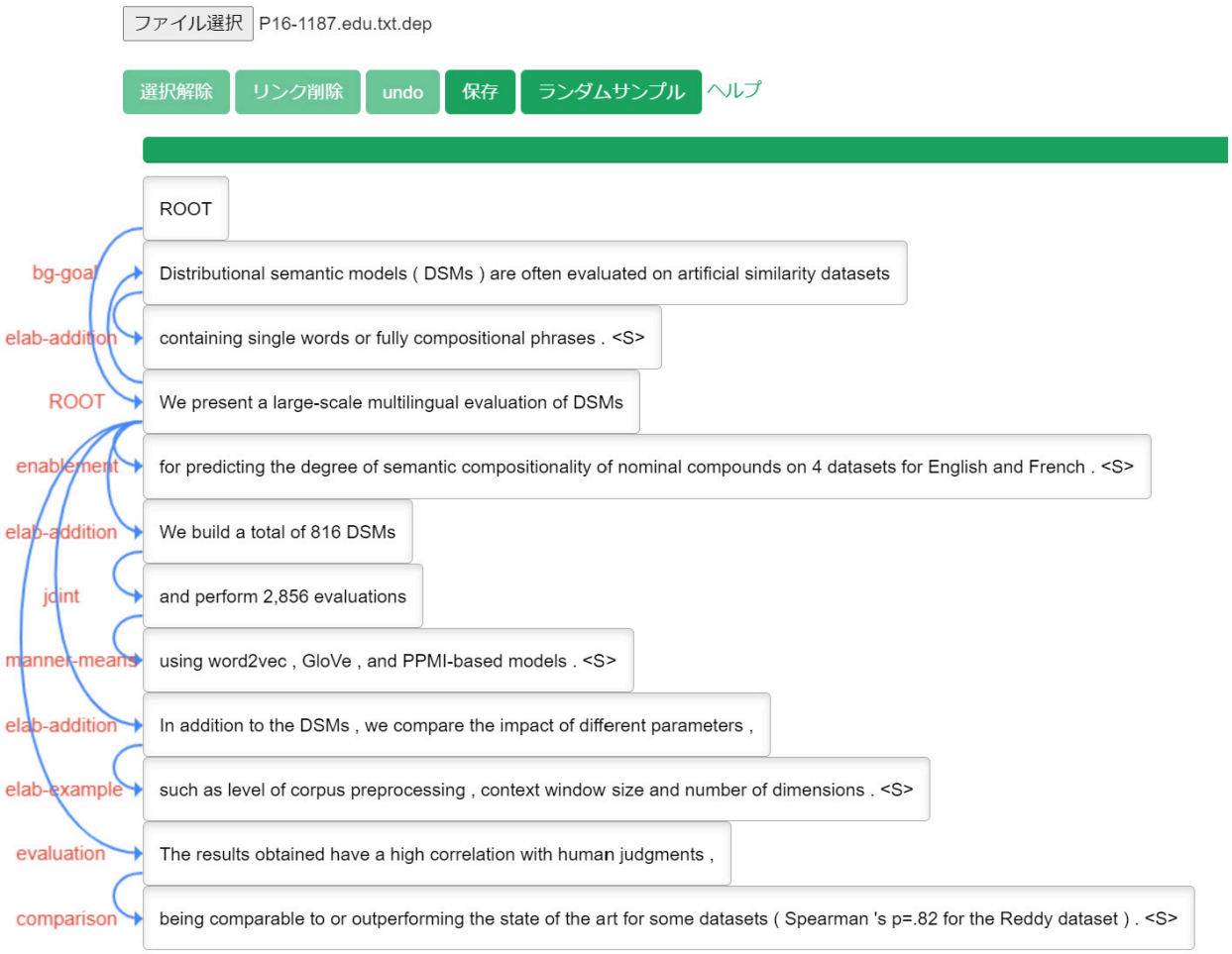


図 1: 論文アブストラクトに対する談話依存構造の例。アノテーションツールによる描画のスクリーンショットを示している。白枠で囲まれた各テキストブロックは談話単位 (Elementary Discourse Unit; EDU) を表し、青い矢印は EDU 間の依存性の有無と方向性を、赤い文字は結合される EDU 間の談話関係を表す。“⟨ S ⟩” は文の切れ目を表す。

4.1 談話単位 (EDU)

談話依存構造では、一つの文章は談話単位 (EDU) に分割される。EDU は、最小単位の連続したテキスト領域であり、EDU 間にオーバーラップはなく、基本的には節 (clause) に対応する。例えば図 1における白枠で囲まれた各テキストブロックはそれぞれ一つの EDU を表しており、先頭には ROOT ノードが付与される。すなわち、文章 d が n 個の EDU に分割されるとき、d は EDU の系列 $d = e_0, e_1, \ldots, e_n$ として表現できる。ここで、e_0 は ROOT ノードに対応する。図 1の例では、1 つのアブストラクトが 11 個の EDU に分割されている。

EDU は基本的に節に対応するが、SciDTB は Carlson ら (2001) のマニュアル [1] に則り、いくつかの例外を設定している。たとえば、文の主語、目的語になる節 (e.g., “*Building a dataset* is important.”) は EDU として分割しない。また、明示的な述語を含まなくても、明示的なディスコースマーカーが付随している名詞句は EDU として分割される (e.g., “[They built a dataset] [*in spite of* the difficulty.]”)。

今回は、既に EDU への分割は人手で行っており、この過程は済んでいるものとする。

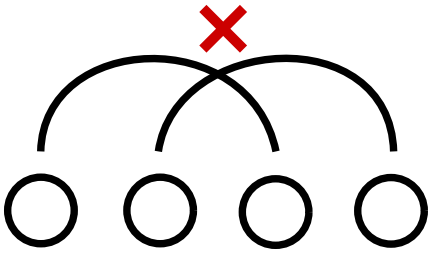


図 2: 交差する依存関係の例。丸マークは EDU を表す。本コーパスでは、交差は起きないものとする。

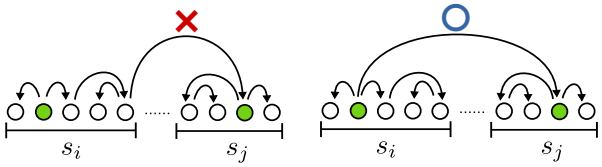


図 3: 文間 (inter-sentence) レベルの談話依存構造では、文全体の中心部 EDU 同士を結合する。丸マークはそれぞれ EDU を表し、文 s_i, s_j それぞれの中心部を緑の丸で表す。

4.2 依存構造

中心部と周辺部 談話依存構造におけるリンクは EDU 間の依存関係の有無と方向性を表す。リンクで結合される 2 つの EDU のうち、中心的な役割を担うほうの EDU は**中心部** (nucleus または head) と呼ばれ、中心部を修飾するほうの EDU は**周辺部** (satellite または dependent) と呼ばれる。図 1のように、**リンクは中心部 (修飾されるほう) から周辺部 (修飾するほう) へと張られる。**

木構造制約 今回、一つの文章 (アブストラクト) には一つの木構造が対応すると考える。これは、**各 EDU は必ず一つだけ修飾先をもち**、かつすべての EDU は連結されている (任意の 2 つの EDU 間にパスが存在する) と定式化することができる。ただし、文章全体の中心部 (必ず一つ) は ROOT ノードを修飾先としてもち、ROOT ノードは修飾先をもたないとする。

非交差制約 今回のコーパスでは、依存関係を表すリンクは**交差しない**という制約を課す。交差する例を図 2に示す。

階層性 談話構造には、文レベルの木構造、段落レベルの木構造、文書レベルの木構造という階層性があることが知られている [2]。そこで、本コーパスの構築でもこの事前知識に基づき、まずは “⟨ S ⟩” 記号をもとに文内 (intra-sentence) レベルで談話依存構造を同定し、そして文間 (inter-sentence) レベルで談話依存構造を同定する。**文間レベルのリンクは、それぞれの文全体の中心部同士を結合することとする。**例えば図 3の (a) では、文 s_i 全体の中心部ではない EDU と文 s_j 全体の中心部が結合するため不適であり、(b) では s_i, s_j それぞれの中心部同士が結合するため適切となる。

Coarse-grained	**Fine-grained**
ROOT	ROOT
ATTRIBUTION (帰属)	attribution (帰属)
BACKGROUND (背景)	bg-compare (背景-比較) bg-general (背景-一般) bg-goal (背景-目標)
CAUSE-EFFECT (原因と結果)	cause (原因) result (結果)
COMPARISON (比較)	comparison (比較)
CONDITION (条件)	condition (条件)
CONTRAST (対比)	contrast (対比)
ELABORATION (拡張)	elab-addition (拡張-追加) elab-aspect (拡張-側面) elab-definition (拡張-定義) elab-enum_member (拡張-要素列挙) elab-example (拡張-例示) elab-process_step (拡張-プロセス)
ENABLEMENT (目的)	enablement (目的)
EVALUATION (評価)	evaluation (評価)
EXPLAIN (説明)	exp-evidence (説明-根拠) exp-reason (説明-理由)
JOINT (並列)	joint (並列)
MANNER-MEANS (方法)	manner-means (方法)
PROGRESSION (追加)	progression (追加)
SAME-UNIT (本来は同じ談話単位)	same-unit (本来は同じ談話単位)
SUMMARY (要約)	summary (要約)
TEMPORAL (時間的)	temporal (時間的)

表 1: 談話関係カテゴリー。coarse-grained では 17 種類、fine-grained では 26 種類が定義されている。今回は、fine-grained の 26 種類のカテゴリーを採用する。

4.3 談話関係

談話関係としては、SciDTB で定義される fine-grained な 26 カテゴリーを採用する。表 1に談話関係カテゴリーを示す。図 1では、各リンクに付与される談話関係は周辺部の EDU の左横に書かれている。

以降は、各談話関係カテゴリーについて例を用いて説明する。

TBA.

5　談話依存構造の例

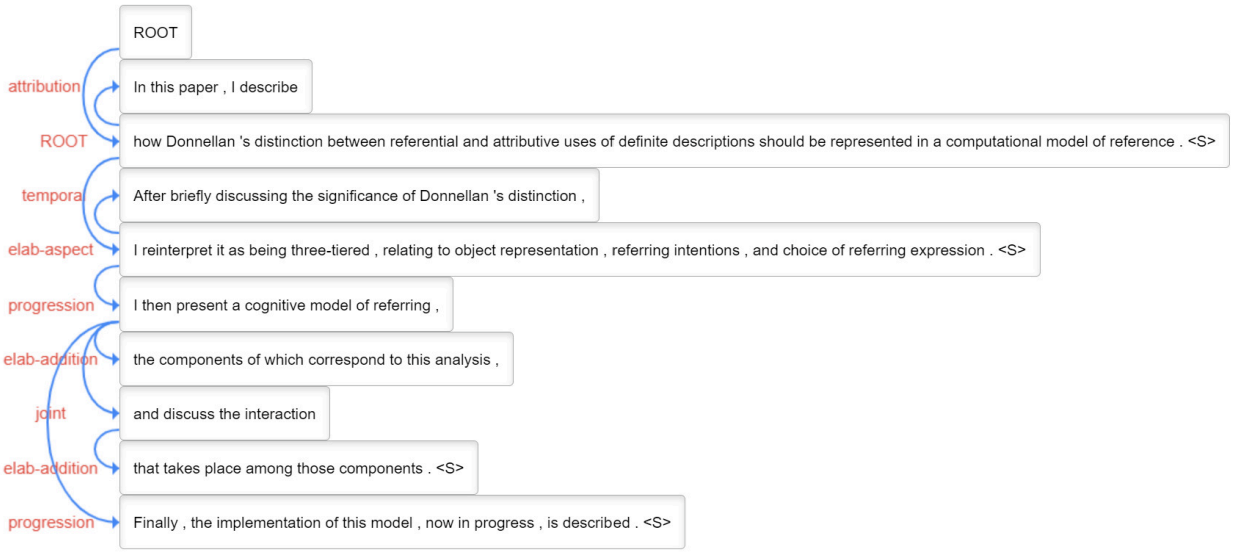
本節では、SciDTB に収録されている自然言語処理論文のアブストラクトに対する談話依存構造の例を示す。より多くの例については、アノテーションツールの「ランダムサンプル」ボタンによって確認することができる。

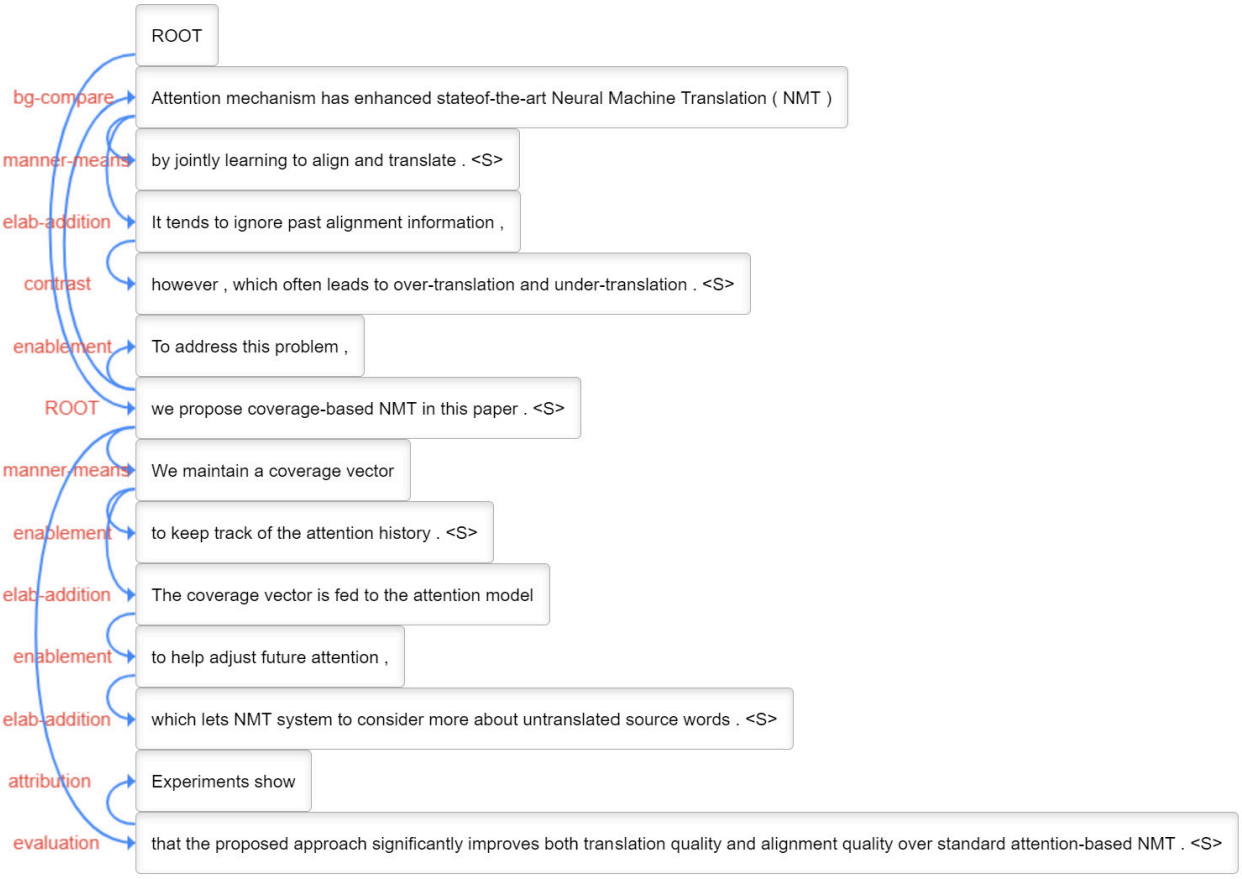
参考文献

[1] Lynn Carlson and Daniel Marcu. Discourse tagging reference manual. In *Technical Report ISI-TR-545*. University California Information Sciences Institute, 2001.

[2] Noriki Nishida and Hideki Nakayama. Unsupervised discourse constituency parsing using Viterbi EM. *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 215–230, 2020.

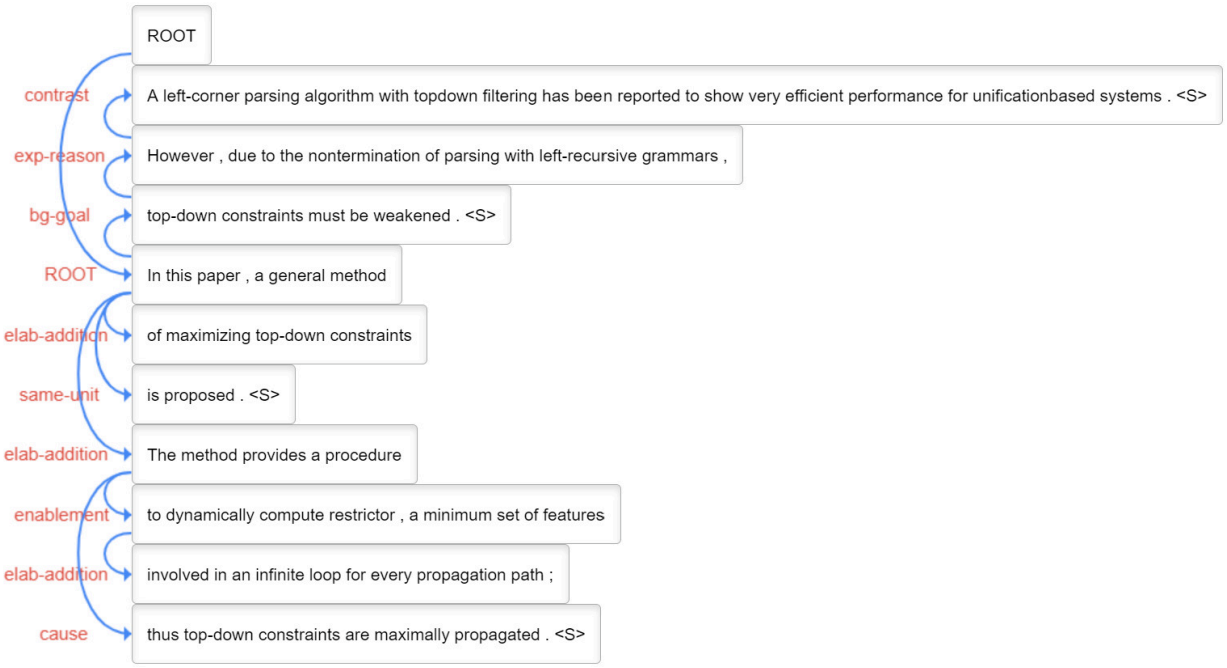
[3] An Yang and Sujian Li. SciDTB: Discourse dependency treebank for scientific abstracts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.


ROOT
attribution
In this paper , I describe
ROOT
how Donnellan 's distinction between referential and attributive uses of definite descriptions should be represented in a computational model of reference . <S>
tempora
After briefly discussing the significance of Donnellan 's distinction ,
elab-aspect
I reinterpret it as being three-tiered , relating to object representation , referring intentions , and choice of referring expression . <S>
progression
I then present a cognitive model of referring ,
elab-addition
the components of which correspond to this analysis ,
joint
and discuss the interaction
elab-addition
that takes place among those components . <S>
progression
Finally , the implementation of this model , now in progress , is described . <S>


ROOT
bg-compare
Attention mechanism has enhanced stateof-the-art Neural Machine Translation (NMT)
manner-means
by jointly learning to align and translate . <S>
elab-addition
It tends to ignore past alignment information ,
contrast
however , which often leads to over-translation and under-translation . <S>
enablement
To address this problem ,
ROOT
we propose coverage-based NMT in this paper . <S>
manner-means
We maintain a coverage vector
enablement
to keep track of the attention history . <S>
elab-addition
The coverage vector is fed to the attention model
enablement
to help adjust future attention ,
elab-addition
which lets NMT system to consider more about untranslated source words . <S>
attribution
Experiments show
evaluation
that the proposed approach significantly improves both translation quality and alignment quality over standard attention-based NMT . <S>


ROOT
ROOT
We present an empirical study on the use of semantic information for Concept Segmentation and Labeling (CSL) ,
elab-addition
which is an important step for semantic parsing . <S>
elab-addition
We represent the alternative analyses output by a state-of-the-art CSL parser with tree structures ,
elab-addition
which we rerank with a classifier
elab-addition
trained on two types of semantic tree kernels :
elab-enum_mer
one processing structures
elab-addition
built with words , concepts and Brown clusters ,
joint
and another one using semantic similarity among the words
elab-addition
composing the structure . <S>
attribution
The results on a corpus from the restaurant domain show
evaluation
that our semantic kernels
elab-addition
exploiting similarity measures
same-unit
outperform state-of-the-art rerankers . <S>


ROOT
bg-goal
Word alignment
elab-addition
using recency-vector based approach
same-unit
has recently become popular . <S>
attribution
One major advantage of these techniques is that
contrast
unlike other approaches
elab-addition
they perform well
contrast
even if the size of the parallel corpora is small . <S>
progression
This makes these algorithms worth-studying for languages
elab-addition
where resources are scarce . <S>
ROOT
In this work we studied the performance of two very popular recency-vector based approaches ,
elab-enum_mem
proposed in (Fung and McKeown , 1994) and (Somers , 1998) , respectively , for word alignment in English-Hindi parallel corpus . <S>
contrast
But performance of the above algorithms was not found to be satisfactory . <S>
contrast
However , subsequent addition of some new constraints improved the performance of the recency-vector based alignment technique significantly for the said corpus . <S>
elab-addition
The present paper discusses the new version of the algorithm and its performance in detail . <S>


ROOT
contrast
A left-corner parsing algorithm with topdown filtering has been reported to show very efficient performance for unificationbased systems . <S>
exp-reason
However , due to the nontermination of parsing with left-recursive grammars ,
bg-goal
top-down constraints must be weakened . <S>
ROOT
In this paper , a general method
elab-addition
of maximizing top-down constraints
same-unit
is proposed . <S>
elab-addition
The method provides a procedure
enablement
to dynamically compute restrictor , a minimum set of features
elab-addition
involved in an infinite loop for every propagation path ;
cause
thus top-down constraints are maximally propagated . <S>


ROOT
bg-goal
Passage-level question answer matching is a challenging task
exp-reason
since it requires effective representations
elab-addition
that capture the complex semantic relations between questions and answers . <S>
ROOT
In this work , we propose a series of deep learning models
enablement
to address passage answer selection . <S>
enablement
To match passage answers to questions
elab-addition
accommodating their complex semantic relations ,
contrast
unlike most previous work
elab-addition
that utilizes a single deep learning structure ,
elab-aspect
we develop hybrid models
elab-addition
that process the text
manner-means
using both convolutional and recurrent neural networks ,
elab-addition
combining the merits on extracting linguistic information from both structures . <S>
elab-aspect
Additionally , we also develop a simple but effective attention mechanism
enablement
for the purpose of constructing better answer representations according to the input question ,
elab-addition
which is imperative for better modeling long answer sequences . <S>
attribution
The results on two public benchmark datasets , InsuranceQA and TREC-QA , show
evaluation
that our proposed models outperform a variety of strong baselines . <S>