医学生物学論文アブストラクトに対する談話依存構造のアノテーション・ガイドライン ver 1.0

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

談話依存構造 (Discourse Dependency Structure) [4, 3] は、文書中の文や節 (clause) の間の修飾・被修飾関係 (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。[1]談話依存構造は、文書が意味的、論理的にどのように構成されているのかを表す。談話依存構造の例を図 1に示す。談話構造は、文書要約や極性分類、情報抽出などで有用であることが示されている。

科学技術論文のアブストラクトの談話依存構造を収録するコーパスとして、SciDTB [5] が存在する。SciDTB は自然言語処理分野の論文アブストラクト 798 件に対して人手で談話依存構造をアノテーションしている。しかし、医学生物学と自然言語処理では論文の書き方や論旨の展開、用いられる語彙の傾向は大きく異なる。例えば、自然言語処理では手法そのものがアブストラクトのメインに置かれる傾向がある一方で、医学生物学では結果や発見内容に重きが置かれる。実際に、私たちは、SciDTB を用いて学習した談話構造解析システムを医学生物学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。[2]

本プロジェクトの目的は、医学生物学の論文アブストラクトに対して人手で談話依存構造を付与したコーパスを構築することである。将来的には、医学生物学論文の談話依存構造の自動解析法およびそこからの医療知識の自動獲得法の開発につなげたい。

2 談話依存構造のアノテーション仕様

談話依存構造のアノテーションは、

1. 談話単位 (Elementary Discourse Unit; EDU) への分割
2. 修飾先 EDU とのリンキング
3. 談話関係のラベリング

の 3 つのステップに分けて行う。尚、2 番目と 3 番目のステップはしばしば同時に行われる。

最初のステップ (EDU 分割) については既に済んでおり、それらは文書ごとに EDU が 1 行ずつ書かれたテキストファイル (*.edu.txt) として保存されている。そのテキストファイルをアノテーションツール

https://norikinishida.github.io/tools/discdep/

にアップロードすると、第 2、第 3 ステップをツール上で行うことができる。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 (*.edu.txt.dep) として保存する。アノテーションツールの使い方については、Section 3で説明する。

[1]談話依存構造は、文書単位のグラフ構造であり、文単位で統語構造を表す依存構造とは異なる。
[2]これは、機械学習では「ドメイン適応」の問題として知られる。

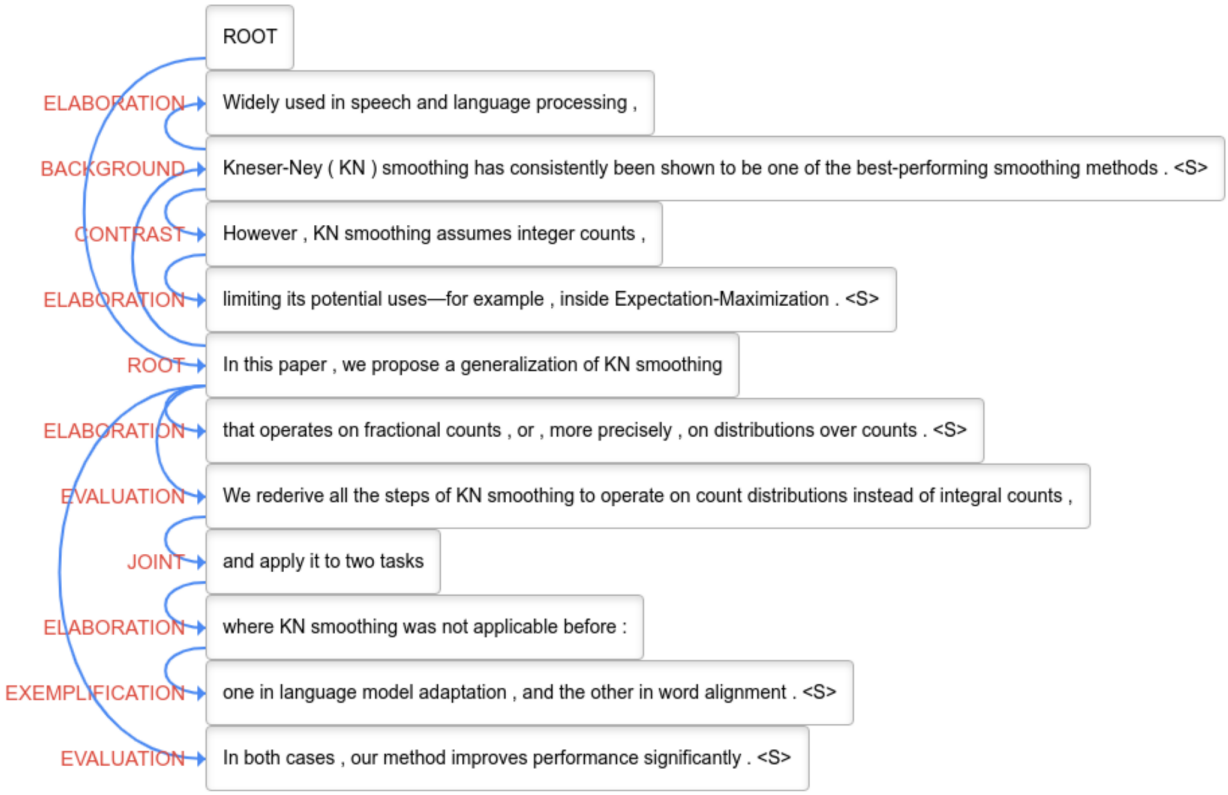


図 1: 論文アブストラクトに対する談話依存構造の例。白枠で囲まれた各テキストブロックは談話単位 (Elementary Discourse Unit; EDU) を表し、青い矢印は EDU 間の修飾関係の有無と方向性を、赤い文字は結合される EDU 間の修飾関係の種類を表す。“⟨ S ⟩” は文の切れ目を表す。

2.1 談話単位 (EDU) への分割

最初のステップでは、アノテーションの対象であるテキストを談話単位 (Elementary Discourse Unit; EDU) に分割する。EDU は、節 (述語) を中心とする最小単位の連続領域であり、EDU 間にオーバーラップはない。例えば、図 1では、アブストラクトは 11 個の EDU に分割されていて、 (白枠で囲まれたテキスト断片がそれぞれ EDU を表している)、文書の先頭には ROOT EDU が付与される。

EDU は基本的に述語のスパンを基準に分割されるが、SciDTB は Carlson ら (2001) のマニュアル [2] に則り、いくつかの例外を設定している。

- 文の main verb の主語、目的語になる述語スパンは EDU としない。

  (1) [Making computers smaller often means sacrificing memory.]

- 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句は EDU として分割する。

  (2) [They went on a picnic] [in spite of the typhoon.]

  (3) [They couldn’t go on a picnic] [due to the typhoon.]

2.2 修飾先 EDU とのリンキング

第 2 ステップでは、各 EDU に対し、それが主に修飾する (係り先の) EDU を同定し、2 つの EDU をリンクで結合する (リンキング)。リンキングの際に注意すべき談話依存構造の制約については、基本的には文の統語的依

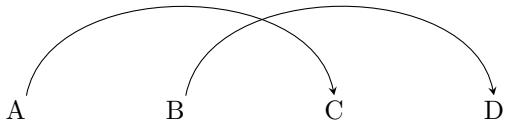


図 2: 交差する談話依存構造の例。A と C のリンクと B と C のリンクが交差している。

	本コーパス	SciDTB
1.	ROOT	ROOT
2.	Attribution (帰属)	Attribution
3.	Background (背景)	Related, Goal, General
4.	Cause-Result-Reason (原因・結果・理由)	Cause, Result, Evidence, Reason
5.	Comparison (比較)	Comparison
6.	Condition (条件・前提)	Condition, Temporal
7.	Contrast (譲歩)	Contrast
8.	Definition (定義)	Definition
9.	Elaboration (詳細化)	Elaboration, Aspect, Process-step, Progression, Summary
10.	Enablement (目的)	Enablement
11.	Evaluation (評価)	Evaluation
12.	Exemplification (例・要素)	Enumerate , Example
13.	Joint (並列)	Joint
14.	Manner-means (方法)	Manner-means
15.	Same-unit (同一 EDU)	Same-unit

表 1: 本コーパスで採用する談話関係カテゴリー。SciDTB における談話関係との対応付けも示す。

存構造の制約と同じである。

中心部と周辺部　談話依存構造におけるリンクは EDU 間の談話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、修飾される側の EDU は**中心部** (被修飾部, head) と呼ばれ、修飾する側の EDU は**周辺部** (修飾部, dependent) と呼ばれる。図 1のように、**リンクは中心部 (被修飾部) から周辺部 (修飾部) へと描く**。

木構造制約　今回、1 つの文章 (アブストラクト) には 1 つの木構造が対応すると考える。これは、各 EDU は必ず 1 つだけ修飾先をもち、かつすべての EDU はグラフ上で連結されている (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、(ROOT を除く) 文書全体の中心部 (必ず一つ) は ROOT EDU をその修飾先としてもち、ROOT EDU 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、リンク同士は**交差しない**という制約を課す。交差する例を図 2に示す。

2.3　談話関係のラベリング

第 3 ステップでは、リンクで結合された EDU 間に対して談話関係ラベルを付与する。談話関係としては、SciDTB および RST Discourse Treebank [2], ISO 24617-8 [1] を参考に 15 種類の談話関係を定義する。表 1に談話関係の一覧を示す。

以降は、15 種類の談話関係カテゴリーについて、それぞれ SciDTB 内の実際のデータ (一部わかりやすさのために修正している) を用いながら説明する。**中心部を太字体で、周辺部を斜字体で示す。**

2.3.1 ROOT

定義 ROOT は、ROOT EDU (中心部) と、文書中で最も重要な EDU (周辺部) との間の関係とする。文書中で最も重要な EDU は、談話依存構造に関する制約上 (Subsection 2.2を参照) 必ず 1 つ存在し、中心部から周辺部への向きにリンクを辿っていくことで他のすべての EDU に到達することができる。したがって、ROOT 関係は必ず 1 回現れる。科学技術論文のアブストラクトでは、“In this paper, ...” や “This study shows” などの表現を伴うことが多い。

例

(4) [**ROOT** $_{\texttt{head}}$] ... [*In this paper, we propose to combine the output from a classification-based system and an SMT-based system* $_{\texttt{dep}}$] ...

(5) [**ROOT** $_{\texttt{head}}$] ... [*This paper presents a negative result on unsupervised domain adaptation for POS tagging.* $_{\texttt{dep}}$] ...

(6) [**ROOT** $_{\texttt{head}}$] ... [*We describe our initial efforts* $_{\texttt{dep}}$] ...

(7) [**ROOT** $_{\texttt{head}}$] ... [*To this end, we propose a novel method* $_{\texttt{dep}}$] ...

2.3.2 Attribution (帰属)

定義 Attribution は、報告内容・主張内容・認知内容を表す中心部と、そのソースを表す周辺部との間の関係とする。基本的には、*attribution verb* と that 節によって明示的に示される。that 節だけではなく、疑問詞で始まる節も対象になる。また、“according to” によってソースが明記される場合は、“according to” で始まる句との間に Attribution 関係を認める。

例

(8) [*X shows* $_{\texttt{dep}}$] [**that Y is Z.** $_{\texttt{head}}$]

(9) [*X argues* $_{\texttt{dep}}$] [**that Y is Z.** $_{\texttt{head}}$]

(10) [*X found* $_{\texttt{dep}}$] [**that Y is Z.** $_{\texttt{head}}$]

(11) [*X noticed* $_{\texttt{dep}}$] [**that Y is Z.** $_{\texttt{head}}$]

(12) [*X investigated* $_{\texttt{dep}}$] [**why Y is Z.** $_{\texttt{head}}$]

(13) [*X investigated* $_{\texttt{dep}}$] [**whether Y is Z.** $_{\texttt{head}}$]

(14) [**Y is Z** $_{\texttt{head}}$] [*according to X.* $_{\texttt{dep}}$]

例外 ソースが明記されてない場合は、Attribution 関係は認めず、EDU 分割もされない。

(15) [It is said that Y is Z.]

2.3.3 Background (背景)

定義 Background は、中心部と、その背景情報を表す周辺部との間の関係とする。科学技術論文ドメインでは、それは研究の動機となった社会的状況であったり、既存研究等の学術的動向であったり、または開発した技術が基づいた基盤技術であったりと、様々である。“based on” などのディスコースマーカーを伴うこともあるが、多くの場合は明示的には示されない。また、他の談話関係に比べ、周辺部が中心部よりも前方に位置することが多い。

例

(16) [*Dependency parsing is a core task in NLP,* $_{\texttt{dep}}$] ... [**We present a new GFL/FUDG-annotated Chinese treebank with more than 18K tokens from Sina Weibo** $_{\texttt{head}}$]

(17) [*Microblog has become a major platform for information about real-world events.* $_{\texttt{dep}}$] ... [**In this study, we focus on the problem of community-related event detection by community emotions.** $_{\texttt{head}}$]

(18) [**We describe a new algorithm for PCFG induction** $_{\texttt{head}}$] [*based on a principled approach* $_{\texttt{dep}}$]

2.3.4 Cause-Result-Reason (原因・結果・理由)

定義 Cause-Result-Reason は、原因と結果、または主張と理由 (根拠) との間の関係とする。原因と結果の場合、どちらが中心部、周辺部になるかは文脈次第であり、文書中でより重要な方を中心部とする。主張と理由 (根拠) の場合、理由 (根拠) 側を常に周辺部とする。“because” や “since”, “as”, “due to”, “because of”, “as a result” などのディスコースマーカーを伴うことが多い。しかし、しばしば語彙情報や統語情報、文脈情報を用いて同定する必要がある。

例

(19) [**This is usually problematic** $_{\texttt{head}}$] [*because lexical ambiguity is ubiquitous,* $_{\texttt{dep}}$]

(20) [**existing models may not do the prediction task well** $_{\texttt{head}}$] [*due to their weakness in sentiment extraction.* $_{\texttt{dep}}$]

(21) [**Chat langiuage is different from natural language** $_{\texttt{head}}$] [*due to its anomalous and dynamic natures,* $_{\texttt{dep}}$] ...

(22) [**Language transfer, the characteristic second language usage patterns** $_{\texttt{head}}$] [*caused by native language interference* $_{\texttt{dep}}$] is investigated ...

(23) [**Parsers typically suffer from the domain mismatch issue,** $_{\texttt{head}}$] [*and thus perform poorly on social media data.* $_{\texttt{dep}}$]

(24) We find [**that the proposed system is robust to disfluencies,** $_{\texttt{head}}$] [*so that a separate stage to elide disfluencies is not required.* $_{\texttt{dep}}$]

(25) [*It only requires consistency between training and testing.* $_{\texttt{dep}}$] [**As a result, there is a wide range of possible preprocessing choices for data** $_{\texttt{head}}$]

(26) We show [**how the expected BLEU objective allows us to train a simple linear discriminative reordering model with millions of sparse features on hundreds of thousands of sentences** $_{\texttt{head}}$] [*resulting in significant improvements.* $_{\texttt{dep}}$]

(27) [**the existing greedy algorithm often selects poor anchor words,** $_{\texttt{head}}$] [*reducing topic quality and interpretability.* $_{\texttt{dep}}$]

(28) [**Our work is novel** $_{\texttt{head}}$] [*in that it explicitly addresses the need* $_{\texttt{dep}}$] ...

(29) [**In general, the recognition problem is undecidable for unification grammars.** $_{\texttt{head}}$] ... [*the problem is computationally hard.* $_{\texttt{dep}}$]

2.3.5 Comparison (比較)

定義 Comparison は、明確に比較される EDU 間の関係とする。比較対象として明示的に挙げられる方を周辺部とする。後述の Contrast (譲歩) との違いに注意。“when compared with” などのディスコースマーカーを伴うことが多い。

例

(30) [*Compared to the standard CCG parser,* $_{\texttt{dep}}$] [**our model is more accurate.** $_{\texttt{head}}$]

(31) [**our ensembles yield significantly better results** $_{\texttt{head}}$] [*when compared with state-of-the-art.* $_{\texttt{dep}}$]

(32) [**It is able to identify utterances with grammatical errors with an F1-score as high as 0.623,** $_{\texttt{head}}$] [*as compared to a baseline F1 of 0.350 on the same data.* $_{\texttt{dep}}$]

2.3.6 Condition (条件・前提)

定義 Condition は、中心部と、その条件や仮定、前提を表す周辺部との間の関係とする。“if” や “when”, “as far as”, “, given that ...’ などのディスコースマーカーを伴うことが多い。

例

(33) [**high-accuracy sentiment analysis is only possible** $_{\texttt{head}}$] [*if word senses with different polarity are accurately recognized.* $_{\texttt{dep}}$]

(34) [**Both methods,** $_{\texttt{head}}$] [*when used with the exponential loss function,* $_{\texttt{dep}}$] bear strong resemblance to the boosting algorithm

(35) [**The method is adaptable to any language,** $_{\texttt{head}}$] [*as far as resources are available.* $_{\texttt{dep}}$]

(36) [*Given a parallel corpus,* $_{\texttt{dep}}$] [**semantic projection attempts to transfer semantic role annotations from one language to another** $_{\texttt{head}}$]

(37) [**The approach is applicable to any type of MWE in any language,** $_{\texttt{head}}$] assuming [*that the MWE is contained in Wikitionary.* $_{\texttt{dep}}$]

(38) [*When the word is ambiguous,* $_{\texttt{dep}}$] [**a disambiguation procedure must be applied.** $_{\texttt{head}}$]

例 (37) では、少しわかりにくいが、“assuming” と “that the MWE is contained in Wikitionary” の間には Attribution 関係があり (“assuming” が周辺部)、“that the MWE...” は “The approach is applicable...” が想定する内容なので、Condition という関係が付けられる。

2.3.7 Contrast (譲歩)

定義 Contrast は、逆説にある EDU 間の関係とする。どちらの EDU が中心部、周辺部になるかは文脈的に重要な方を中心部とする。“however” や “although”, “while”, “instead”, “in spite of” などのディスコースマーカーを伴うことが多いが、文脈的に判断すべきケースもある。

例

(39) [**Knowledge graphs are recently used,** head] ... [*However, few of the methods pay attention to non-entity words* dep]

(40) [**There is rising interest in vector-space word embeddings and their use in NLP,** head] ... [*Nealy all this work, however, assumes a single vector per word type* dep]

(41) [*Although the training objective is no longer concave,* dep] [**it can still be used to improve an initial model** head]

(42) [*while labeled data in NLP is heavily biased* dep], [**importance weighting has seen only few applications in NLP,** head]

(43) [**We do not trust the best or any specific query segmentation.** head] [*Instead, evidence in favor of candidate e2e are aggregated across several segmentations.* dep]

(44) [*Rather than finding an approximate convex hull in a high-dimensional word co-occurrence space,* dep] [**we propose to find an exact convex hull in a visualizable 2- or 3-dimensional space.** head]

2.3.8 Definition (定義)

定義 Definition は、中心部と、その定義や言い換えを表す周辺部との間の関係とする。しばしば括弧やコロン、"i.e.", "that is," 等のマーカーを伴うが、それらのマーカーが付随しても厳密には定義、言い換えではない場合 (後述の Exempliification) もあるため、文脈的に判断する必要がある。

例

(45) [**When the word is ambiguous** head] [*(there are several possible analyses for the word),* dep] ...

(46) [**Context-predicting models** head] [*(more commonly known as embeddings or neural language models)* dep] are ...

(47) [**This process models exploratory search:** head] [*a user explores a new topic* dep] ...

(48) [**The major NLP challenge for personal assistants is machine understanding:** head] [*translating natural language user commands into an executable representation.* dep]

(49) [**We apply this framework to the task of Semantic Textual Similarity (STS)** head] [*(i.e., judging the semantic similarity of natural-language sentences).* dep]

(50) [**We present WiBi, an approach to the automatic creation of a bitaxonomy for Wikipedia,** head] [*that is, an integrated taxonomy of Wikipedia pages and categories.* dep]

2.3.9 Elaboration (詳細化)

定義 Elaboration、既出の中心部と、その中心部に対してさらに詳細な情報を付加する周辺部との間の関係とする。Elaboration。他の談話関係のいずれも該当しない場合は、Elaboration。

文内の Elaboration については、構文情報が有用な指標になり、例えば中心部と、それが含む名詞句を関係節、前置詞句の形で後ろから修飾する周辺部との間の関係は Elaboration とする。文間の Elaboration については、"moreover", "furthermore" 等のディスコースマーカーは存在するが、多くの場合は明示的ではなく、共参照関係や文脈情報を用いて同定する必要がある。

例

(51) [**We introduce a new CCG parsing model** $_{\texttt{head}}$] [*which is factored on lexical category assignments.* $_{\texttt{dep}}$]

(52) [**Language identification and transliteration for Hindi are two major challenges** $_{\texttt{head}}$] [*that impact POS tagging accuracy.* $_{\texttt{dep}}$]

(53) [**We investigate the possibility** $_{\texttt{head}}$] [*to automatically generate sports news from live text commentary scripts.* $_{\texttt{dep}}$]

(54) [**Knowledge takes the form of lexicalized assertions** $_{\texttt{head}}$] [*associated with open-domain classes.* $_{\texttt{dep}}$]

(55) [**The dynamic reranking model achieves an absolute 1.78% accuracy improvement over the deterministic baseline parser on PTB,** $_{\texttt{head}}$] [*which is the highest improvement by neural rerankers in the literature.* $_{\texttt{dep}}$]

(56) [**We introduce a new CCG parsing model** $_{\texttt{head}}$] which is factored on lexical category assignments. [*Parsing is then simply a deterministic search for the most probable category sequence* $_{\texttt{dep}}$]

(57) [**Different approaches to high-quality grammatical error correction have been proposed recently.** $_{\texttt{head}}$] [*Most of these approaches are based on classification or statistical machine translation.* $_{\texttt{dep}}$]

(58) [**Importance weighting is a generalization of various statistical bias correction techniques.** $_{\texttt{head}}$] [*Importance weighting has seen only few applications in NLP.* $_{\texttt{dep}}$]

(59) [**The matrices perform better than full tensors,** $_{\texttt{head}}$] [*allowing a reduction in the number of parameters* $_{\texttt{dep}}$]

(60) [**In this paper, we propose a novel question difficulty estimation approach** $_{\texttt{head}}$] ... [*We further employ a K-Nearest approach* $_{\texttt{dep}}$]

(61) [**We present a new GFL/FUDG-annotated Chinese treebank with more than 18K tokens from Sina Weibo.** $_{\texttt{head}}$] [*We formulate the dependency parsing problem as many small and parallelizable arc prediction tasks.* $_{\texttt{dep}}$]

(62) [**We propose a new Chinese abbreviation prediction method** $_{\texttt{head}}$] ... [*We introduce the minimum semantic unit* $_{\texttt{dep}}$] ... [*We use an integer linear programming (ILP) formulation with various constraints* $_{\texttt{dep}}$]

(63) [**We present a novel translation model** $_{\texttt{head}}$] ... [*A tree-to-string alignment template is capable of generating both terminals and non-terminals* $_{\texttt{dep}}$] ... [*The model is linguistically syntax-based* $_{\texttt{dep}}$] ...

(64) [**In this method, the dependency parsing is executed in two stages: at the clause level and the sentence level.** $_{\texttt{head}}$] [*First, the dependencies within a clause are identified* $_{\texttt{dep}}$] ... [*Next, the dependencies over clause boundaries are identified stochastically,* $_{\texttt{dep}}$] ...

(65) [**This paper proposed a method** $_{\texttt{head}}$] ... [*First, it learns decision lists from training data* $_{\texttt{dep}}$] ... [*Then, it is augmented by feedback* $_{\texttt{dep}}$] ... [*Finally, it detects errors* $_{\texttt{dep}}$] ...

(66) [**we: (i) leverage content from the local neighborhood of a user;** $_{\texttt{head}}$] [*(ii) evaluate batch models as a function of size and the amount of messages in various types of neighborhood;* $_{\texttt{dep}}$] [*and (iii) estimate the amount of time and tweets* $_{\texttt{dep}}$]

(67) [**We describe a new algorithm for PCFG induction** $_{\texttt{head}}$] ... [*Moreover, this algorithm can work on large grammars and datasets* $_{\texttt{dep}}$]

2.3.10 Enablement (目的)

定義 Enablement は、中心部と、その目的を表す周辺部との間の関係とする。“in order to” や “so as to” などのディスコースマーカーや、「目的」を表す to 不定詞や for + 現在分詞によって明示的に示されることが多い。

例

(68) [*In order to capture more keywords,* $_{\texttt{dep}}$] [**we also incorporate syntactic information into the CBOW model.** $_{\texttt{head}}$]

(69) [*To address this task,* $_{\texttt{dep}}$] [**we exploit extra-textual information** $_{\texttt{head}}$] ...

(70) [**The system can be used** $_{\texttt{head}}$] [*to aid BIO-NLP directory or as useful material* $_{\texttt{dep}}$]

(71) [**A number of lexical association measures have been studied** $_{\texttt{head}}$] [*to help extract new scientific terminology or general-language collocations.* $_{\texttt{dep}}$]

(72) [**In this paper we propose a method** $_{\texttt{head}}$] [*to increase dependency parser performance* $_{\texttt{dep}}$]

(73) [**This paper proposes to apply the continuous vector representations of words** $_{\texttt{head}}$] [*for discovering keywords from a financial sentiment lexicon.* $_{\texttt{dep}}$]

(74) [**We present a weakly-supervised algorithm** $_{\texttt{head}}$] [*for harvesting semantic relations.* $_{\texttt{dep}}$]

2.3.11 Evaluation (評価)

定義 科学技術論文ドメインにおいて Evaluation は、中心部と、その実験結果、評価結果を表す周辺部との間の関係とする。特有のディスコースマーカーは存在せず、文脈的に判断する必要がある。

例

(75) [**We propose a neural network approach** $_{\texttt{head}}$] ... [*the proposed method marks new state-of-the-art accuracies for English POS tagging tasks.* $_{\texttt{dep}}$]

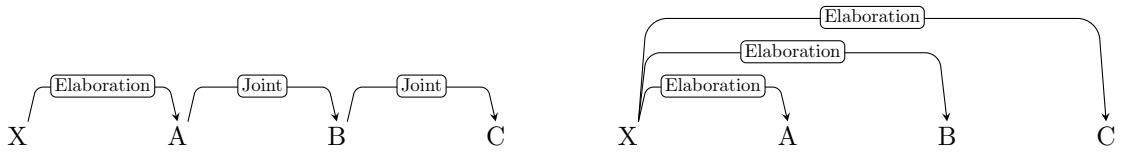
(76) [**In this paper, we propose to combine the output from a classification-based system and an SMT-based system** $_{\texttt{head}}$] ... [*We achieve an F0.5 score of 39.39% on the test set of the CoNLL-2014 shared task* $_{\texttt{dep}}$]

(77) [**We investigate grammatical error detection in spoken language,** $_{\texttt{head}}$] ... [*The proposed system outperforms two baseline systems on two different corpora* $_{\texttt{dep}}$]

2.3.12 Exemplification (例・要素)

類似カテゴリー Definition (定義)

定義 Exemplification は、中心部と、その具体例や要素 (項目) を最低 1 個以上列挙する周辺部との間の関係とする。“such as” 等のディスコースマーカーや、Definition 同様に括弧やコロン、“i.e.”, “e.g.” 等のマーカーを伴うことが多い。また、“first”, “second” や “(1)”, “(2)” のような明示的なリスティングマーカーを伴うこともある。



(a) Joint を用いる場合 (本コーパス)　　(b) Joint を用いない場合

図 3: Joint 関係のアノテーション方針。EDU X に対して EDU A, B, C が同じ談話関係 (e.g., Elaboration) で結合し、かつ A, B, C が “and” 等を用いて並列関係である場合、(b) の代わりに (a) のようにアノテーションする。

例

(78) [**Existing methods only employ the internal translation similarity** head] [*such as content-based similarity and page structural similarity* dep]

(79) [**It resolves the ambiguities in the main three steps of QALD** head] [*(phrase detection, phrase-to-semantic-item mapping, and semantic item grouping).* dep]

(80) [**This paper proposes a generic mathematical formalism for the combination of various structures:** head] [*strings, trees, dags, graphs and products of them.* dep]

(81) [**We analyze the importance of seed lexicons for the SBWES induction across different dimensions** head] [*(i.e., lexicon source, lexicon size, translation method, translation pair reliability).* dep]

(82) [**Wikification links each concept mention to a concept referent in a knowledge base** head] [*(e.g., Wikipedia).* dep]

(83) [**Here we address two limitations of this approach** head] ... [*First, Web queries are rarely well-formed questions.* dep] ... [*Second, the knowledge graph is always incomplete,* dep] ...

2.3.13 Joint (並列)

定義　Joint は、2 つ以上の EDU が “and” 等のディスコースマーカーによって並列する場合の関係とする。それらの EDU の重要性は等しいが、従来法に従ってひとつ左側の EDU を中心部、その一つ右側を周辺部として、チェインしていく。例を図 3に示す。例えば、X (EDU) の詳細を A, B, C の 3 つの EDU によって説明するとき (Elaboration 関係)、Elaboration(X, A), Joint(A, B), Joint(B, C) の 3 つの談話関係をアノテーションする。図 3の右図のように、Joint を用いずに、Elaboration(X, A), Elaboration(X, B), Elaboration(X, C) というようにすることもできるが、アノテーションの一貫性を保つためと、並列関係を捉えるために、Joint を用いて前者のようにアノテーションする。

例

(84) [**We propose a probabilistic model for POS guessing of unknown words** head] [*and estimate its parameters.* dep]

(85) [**We attempt to apply the model to semisupervised learning,** head] [*and conduct experiments on multiple corpora.* dep]

(86) As a solution we describe a method [**enlarging the vocabulary of a language model to an almost infinite size** head] [*and capturing their context information* dep]

(87) [**MAGEAD performs an online analysis to or generation from a root+pattern+features representation,** head] [*it has separate phonological and orthographic representations,* dep] [*and it allows for combining morphemes from different dialects.* dep]

2.3.14 Manner-means (手段)

定義 Manner-means は、中心部と、そのための手段を表す周辺部との間の関係とする。“using ...” や「手段」の “by” などのディスコースマーカーを伴うことが多い。

例

(88) [**We propagate the enriched features in a graph** $_{\texttt{head}}$] [*using an unsupervised algorithm.* $_{\texttt{dep}}$]

(89) [**We evaluate these models** $_{\texttt{head}}$] [*using customer agent dialogs from a catalog service domain.* $_{\texttt{dep}}$]

(90) [**An experiment** $_{\texttt{head}}$] [*using a spoken monologue corpus* $_{\texttt{dep}}$] shows ...

(91) [*By incorporating textual information,* $_{\texttt{dep}}$] [**RCM can effectively deal with data sparseness problem.** $_{\texttt{head}}$]

(92) [**Existing methods incrementally expand the lexicon** $_{\texttt{head}}$] [*by greedily adding entities.* $_{\texttt{dep}}$]

(93) [*Through a simple bootstrapping procedure,* $_{\texttt{dep}}$] [**we learn the likelihood of coreference between a pronoun and a candidate noun** $_{\texttt{head}}$]

2.3.15 Same-unit (同一 EDU)

定義 他の談話関係とは異なり、Same-unit はダミー的な関係カテゴリーである。名詞句を後ろから修飾する節が EDU として認められる都合上、もともとは 1 つの EDU が、そのような EDU の埋め込みによって 2 つの EDU に分割されてしまうことがある。例えば、“Charles Lutwidge Dodgson, better known by his pen name Lewis Carroll, was an English writer of children’s fiction.” という文は、以下のように A, B, C の 3 つの EDU に分割される。

(94) [Charles Lutwidge Dodgson,]$_{\text{A}}$ [better known by his pen name Lewis Carroll,]$_{\text{B}}$ [was an English writer of children’s fiction.]$_{\text{C}}$

これは、B (“better known by ...”) が 1 つの EDU をなすからであり、もし B がなければ A と C は 1 つの EDU (“Charles Lutwidge Dodgson was an English writer of children’s fiction.”) であり、A と C の間に談話関係はない。Same-unit は、これらのもともとは同一の EDU に属するが分割されてしまった EDU を結合するときに使う。文よりも大きい EDU は存在しないため、Same-unit は同一文内の EDU 間でのみ生じる。左側の EDU を中心部とする。

例

(95) [**the only manual annotations** $_{\texttt{head}}$] needed for training [*are grammatical error labels.* $_{\texttt{dep}}$]

(96) [**The experimental results** $_{\texttt{head}}$] using open benchmarks [*demonstrate the effectiveness of the proposed method.* $_{\texttt{dep}}$]

(97) [**A vote prediction system** $_{\texttt{head}}$] that exploits only textual information [*can be improved significantly* $_{\texttt{dep}}$]

3 アノテーションツールの使い方

本節では、アノテーションツールの使い方について説明する。

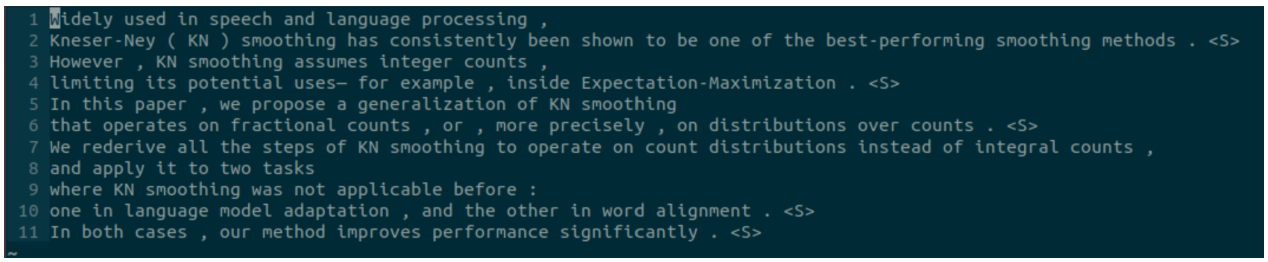


図 4: EDU データ (*.edu.txt) の例。

3.1 準備

- アノテーションツール `https://norikinishida.github.io/tools/discdep/`
- EDU データ (*.edu.txt) 200 件

アノテーションツールは上記 URL でアクセスすることができる。特別なインストールは必要ない。アノテーション環境としては Google Chrome を想定している (が、他のブラウザでも動くかもしれない)。EDU データは、1 行ごとに 1EDU が記述された文書 (アブストラクト) ごとのテキストファイルである。EDU データの中身の例を図 4に示す。

3.2 EDU データのアップロード

初期画面では、図 5の上図のように、「ファイル選択」ボタンと「ランダムサンプル」ボタン、本ガイドラインへのリングのみが表示される。「ランダムサンプル」ボタンを押すことで、SciDTB 内のアノテーションデータをランダムに確認することができる。

「ファイル選択」ボタンを押し、EDU データをアップロードすると、図 5の下のように、1 行ごとに記述されていた EDU が展開される。「選択解除」ボタン、「リンク解除ボタン」、「undo」ボタン、「保存」ボタンはそれぞれまだ押せない。ボタン群の下のバーはプログレスバーであり、これが 100%になると「保存」ボタンは押せるようになる。

3.3 談話関係のアノテーション

談話関係のアノテーションでは、まず談話関係の中心部 (被修飾部) となる EDU を左クリックで選択する (図 6の上図)。次に、中心部が選択された状態で、談話関係の周辺部 (修飾部) となる EDU を同様に選択する。すると、それらの間の談話関係カテゴリーを選択するためのダイアログが出てくるので (図 6の真ん中)、談話関係カテゴリーを選択し、ダイアログの「OK」ボタンを押す。そうすると、図 6の下図のように談話関係が追加される。

他の談話関係を追加するには、同様に「中心部選択 → 周辺部選択 → 談話関係選択」を繰り返せばよい。

3.4 選択解除、リンク削除、undo

アノテーションの過程で、EDU の選択を取り消したい場合や、既にアノテーションした談話関係リンクを削除したい場合がある。選択を解除するには、解除しい EDU を選択した状態で、「選択解除」ボタンを押す。談話関係リンクを削除したい場合は、そのリンクの**周辺部**を選択し、その状態で「リンク削除」ボタンを押す。EDU の選択や解除、談話関係の追加や削除をやり直したい場合は、「undo」ボタンを押す。

3.5 保存

すべての談話関係を追加すると、プログレスバーは 100%になり、「保存」ボタンを押せるようになる。JSON ファイル (*.edu.txt.dep) が出力されるので、それを保存する。図 7に出力される JSON ファイルの中身の例を示す。

参考文献

[1] Harry Bunt and Rashmi Prasad. ISO DR-Core (ISO 24617-8): Core concepts for the annotation of discourse relations. In *Proceedings of 12th Joint ACL-ISO Workshoop on Interoperable Sementic Annotation (ISA-12)*, 2016.

[2] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*, 2001.

[3] William C. Mann and Sandra A. Thompson. Rhetorical Structure Theory: Towards a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, Vol. 8, No. 3, pp. 243–281, 1988.

[4] Mathieu Morey, Philippe Muller, and Nicholas Asher. A dependency perspective on RST discourse parsing and evaluation. *Computational Linguistics*, Vol. 44, No. 2, pp. 197–235, 2018.

[5] 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.

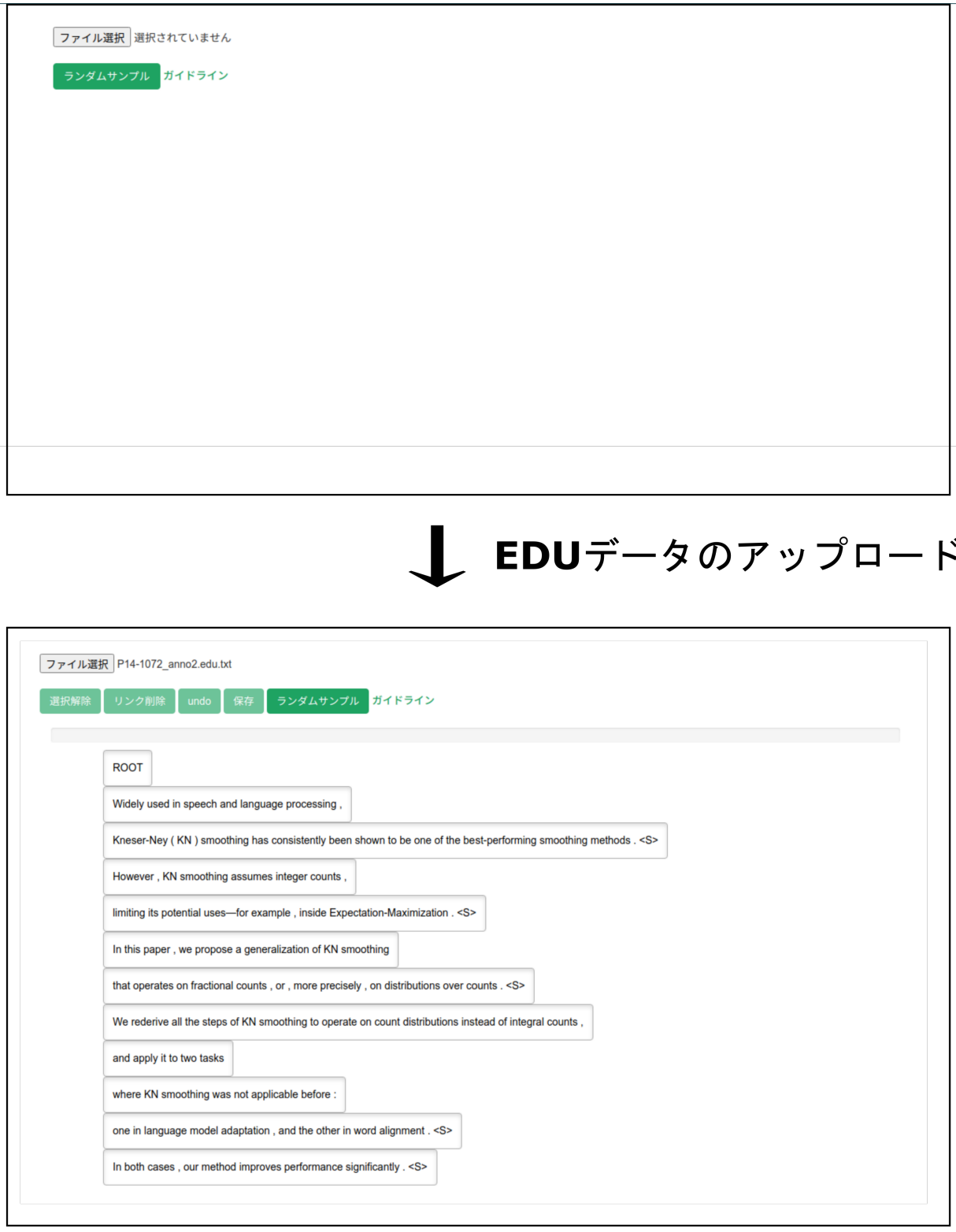
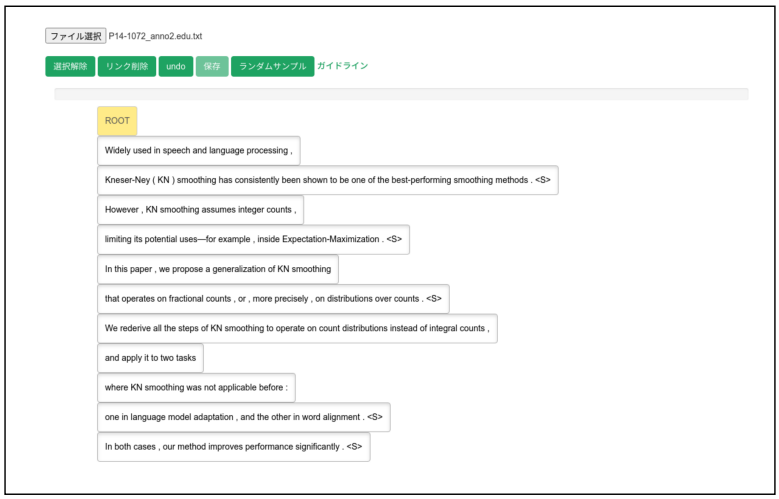
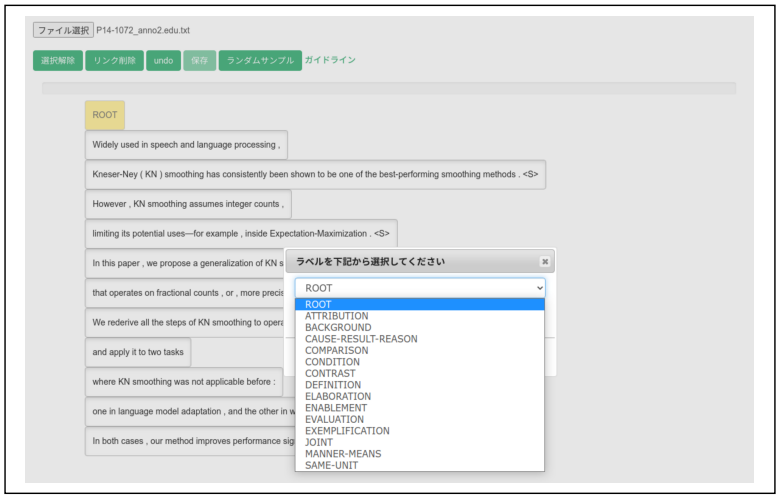


図 5: EDU データをアップロードしたときの画面。EDU が展開される。



周辺部を選択



談話関係を選択

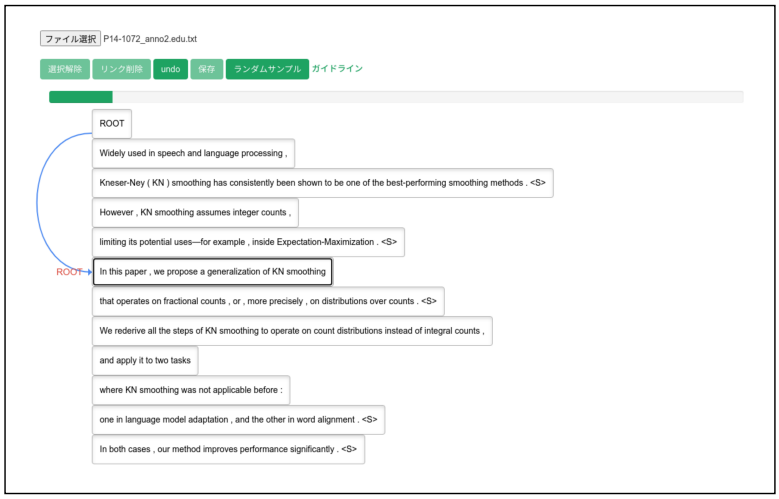
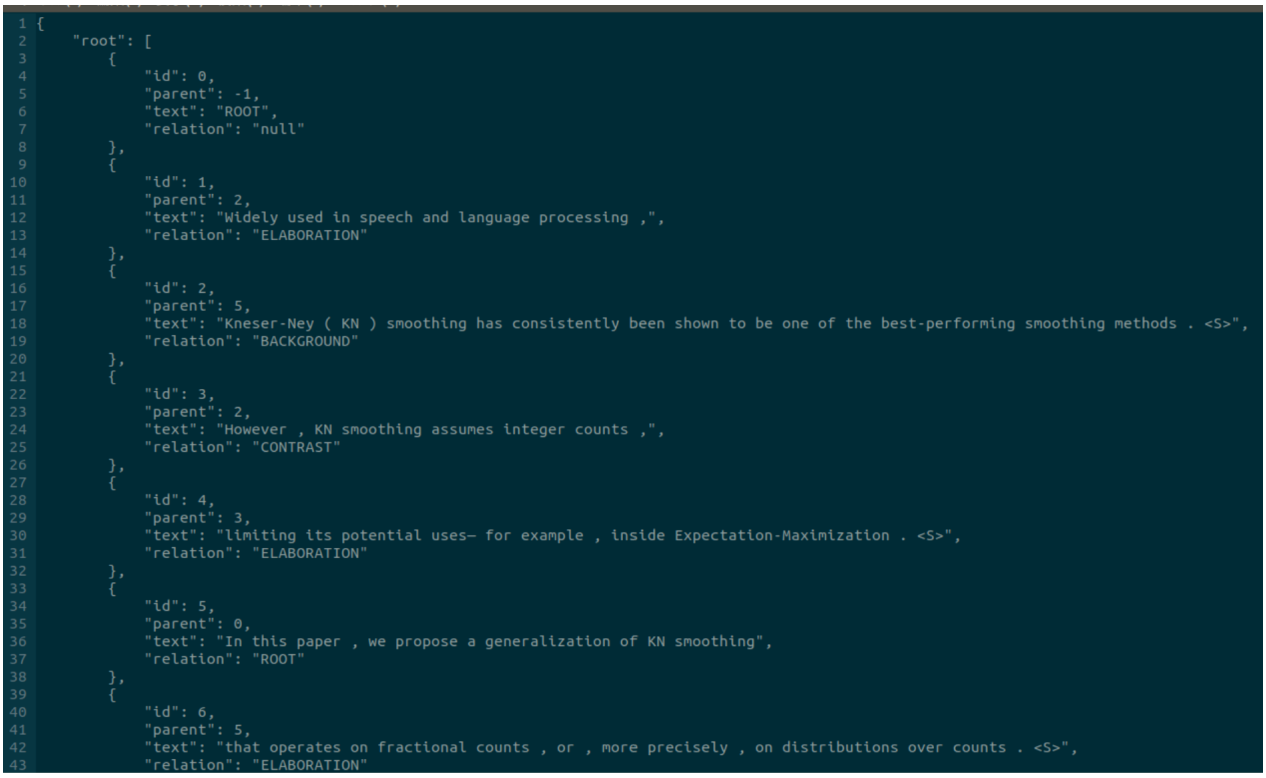


図 6: 談話関係の追加。ここでは ROOT EDU と 6 番目の EDU (“In this paper, ...”) の間に ROOT 関係を追加している。



```
{
    "root": [
        {
            "id": 0,
            "parent": -1,
            "text": "ROOT",
            "relation": "null"
        },
        {
            "id": 1,
            "parent": 2,
            "text": "Widely used in speech and language processing ,",
            "relation": "ELABORATION"
        },
        {
            "id": 2,
            "parent": 5,
            "text": "Kneser-Ney ( KN ) smoothing has consistently been shown to be one of the best-performing smoothing methods . <S>",
            "relation": "BACKGROUND"
        },
        {
            "id": 3,
            "parent": 2,
            "text": "However , KN smoothing assumes integer counts ,",
            "relation": "CONTRAST"
        },
        {
            "id": 4,
            "parent": 3,
            "text": "limiting its potential uses— for example , inside Expectation-Maximization . <S>",
            "relation": "ELABORATION"
        },
        {
            "id": 5,
            "parent": 0,
            "text": "In this paper , we propose a generalization of KN smoothing",
            "relation": "ROOT"
        },
        {
            "id": 6,
            "parent": 5,
            "text": "that operates on fractional counts , or , more precisely , on distributions over counts . <S>",
            "relation": "ELABORATION"
```

図 7: 出力される JSON ファイルの中身の例。