情報抽出技術を用いた JCoLA の拡張に向けて

染谷 大河 [1]　進藤 裕之 [2,3]　大関 洋平 [1]
[1] 東京大学　[2] 奈良先端科学技術大学院大学
[3]MatBrain 株式会社
{taiga98-0809, oseki}@g.ecc.u-tokyo.ac.jp
shindo@is.naist.jp

概要

近年多くの場面で活躍するニューラル言語モデルが、自然言語に内在する統語構造をどれほど理解しているのかを検証する研究が盛んである。しかし、検証のためのデータセットを言語学論文の例文をもとに作成する際には、人手での例文抽出が不可欠であり、それゆえにデータセットの規模を容易に拡大することができないという課題があった。本研究では、情報抽出の技術を言語学の分野に応用し、理論言語学の論文や教科書等から自動で例文を抽出することで、複雑な統語現象を扱いつつ、かつ大規模であるという特徴を併せ持つ初めてのデータセットの構築を試みる。

1　はじめに

近年、翻訳や文章要約、文法誤り訂正等さまざまな場面で、ニューラル言語モデルが活躍している。特に Transformer [1] をベースとしたモデルの精度は非常に高く、様々なタスクで高い精度を発揮している [2, 3]。一方、理論言語学の分野では、自然言語には一種の統語構造が存在しているということが伝統的に主張されてきている [4, 5]。確かに、先述のようなニューラル言語モデルは多くの自然言語処理タスクで高性能を発揮し、多くのアプリケーションの基盤となっているが、そのようなニューラル言語モデルが自然言語の統語構造をどれほど理解しているのかについては、未だ多くが分かっていない。

このような背景から、ニューラル言語モデルが実際にどれほどの統語知識を獲得しているのかを検証する研究が盛んに行われている。近年は、特に言語モデルの統語的評価を行うための大規模なデータセットを作った上で、それを用いて言語モデルを評価する研究が盛んになってきている [6, 7, 8, 9]。そのような研究の例としては、Warstadt et al. (2019) [6] や Trotta et al. (2021) [8] のように、より複雑な統語現象を取り扱う理論言語学のジャーナル論文や教科書から例文を抽出することによりデータセットを作成するという方法がある。これにより、自動生成により作成されたデータセットに比べて、より複雑な統語現象を扱うデータセットを作成することができる一方で、例文の抽出作業は人手で行う必要があるため、自動生成の場合と比べてそのデータセットの規模に課題があった。また、このようなデータセットを日本語で作成した例がなかったことなどから、理論言語学のジャーナル論文をもとに JCoLA [10] が構築されたが、依然データセットの規模の問題は解決されていない。

そこで本論文では、情報抽出の技術を言語学の分野に応用し、理論言語学の論文や教科書等から自動で例文を抽出する試みを紹介する。これにより、理論言語学で議論されている複雑な統語現象を扱いつつ、かつ大規模であるという特徴を併せ持つ初めてのデータセットの構築が可能になる。

2　言語モデルの統語的評価

近年は、Linzen et al. (2016) [11] を端緒として、統語的評価用のデータセットを用いて、言語モデルがどれほどの統語知識を獲得しているのかを検証する研究が盛んである。Warstadt et al. (2019) [6] は、理論言語学のジャーナル論文や教科書から例文を抽出し、言語モデルの 2 値分類性能をテストする大規模データセットである CoLA (Corpus of Linguistic Acceptability) [1)] を構築した。CoLA などの言語学の論文中の例文から構築されているデータセット [6, 8] は、理論言語学の論文や教科書から例文を抽出しているため、理論言語学で扱われている複雑な統語現象 (Class Ⅲ judgments, Marantz

1)　英語を対象とした大規模な言語理解ベンチマークである GLUE [2] に含まれるデータセットの一つともなっている。

表 1　既存の統語的評価用大規模データセットと JCoLA

データセット	論文から抽出	ミニマルペア	例文の自動生成	データサイズ
CoLA [6]	✓			10,657 文
ItaCoLA [8]	✓			9,722 文
BLiMP [7]		✓	✓	67,000 ペア
CLiMP [9]		✓	✓	16,000 ペア
JCoLA（本研究）	✓	✓	✓	369 ペア + 約 10,000 文（予定）

2005 [12]; Linzen and Oseki 2018 [13] 参照) を対象とした評価が可能になっているものの、例文の抽出を人手で行っているため、その規模を容易に拡大していくことが困難であるという特徴があった。一方、Warstadt et al. (2020) [7] は、ミニマルペアを自動生成してまとめた大規模データセット BLiMP（The Benchmark of Linguistic Minimal Pairs for English) を構築した。BLiMP など例文の自動生成を行うことで構築されたデータセット [14, 7, 15] は、自動生成をおこなっているためそのデータセットの規模が大きく、さらにその拡大も容易であるが、ある特定のパターンにより例文が生成されるため、複雑な統語現象を扱うことができないという問題点があった。

一方で、以上の言語モデルの統語知識を検証する試みは、その大多数が英語を対象としたものである。一部の研究で対象を英語以外にも拡張した検証 [16, 17, 18, 15] が行われてはいるが、幅広い統語現象を対象とし、かつ英語以外で検証を行った研究は非常に限られている [8, 9]。特に日本語においては自然言語処理の分野で広く使われる言語モデルの統語知識を評価するベンチマークとなるようなデータセットは存在しておらず、言語モデルが英語等の一部の言語だけではなく、自然言語一般の統語現象を捉えられているかどうかについての明確な証拠を得ることはできない状況であると考えられる。

そのような背景から、染谷・大関 (2022) [10] では理論言語学のジャーナル論文から例文を手動で抽出することで、日本語を対象とした言語モデルの統語的評価用データセット JCoLA (Japanese Corpus of Linguistic Acceptability) を作成した。しかし、その作業の性質上データセットの規模の拡張が難しいという問題点は解決することはできず、総例文数は 2,323 文にとどまった（表 1)。

本論文では、すでに化学分野等の論文から自動で図表等の情報を抽出するために用いられている情報抽出の技術を言語学の分野に応用し、理論言語学の論文から自動で例文を抽出する試みを紹介する。これにより、理論言語学で議論されている複雑な統語現象を扱いつつ、かつ大規模であるという特徴を併せ持つ初めてのデータセットの構築が可能となる。

3　JCoLA の拡張

3.1　データ収集

本研究では、東アジア・東南アジア言語の理論言語学のジャーナルとして著名な JEAL (Journal of East Asian Linguistics) 、また世界的に著名な理論言語学のジャーナルである LI (Linguistic Inquiry) と NLLT (Natural Language Linguistic Theory) のそれぞれにおいて 2006 年から 2020 年の 15 年間で掲載された論文の中で、特に日本語の統語論を扱う論文を対象として、それらの論文に含まれる例文の自動抽出を行う。

これにより、理論言語学の分野で頻繁に議論されているような、複雑な統語現象 (Class Ⅲ judgments, Marantz 2005[12]; Linzen and Oseki 2018[13] 参照) を伴う例文を収集することができると考えられる。また、レイアウトの異なる複数のジャーナルを対象とすることで、各ジャーナルごとのレイアウトの差異に堅牢な情報抽出モデルの学習が可能となる。

3.2　訓練・評価用データの作成

前節で収集した論文 PDF はそのままでは情報抽出モデルの訓練・評価用データとして直接的に用いることは出来ず、情報抽出モデルの入力として適切な形式にするには、PDF からテキストを抽出し、かつどの範囲が抽出すべき例文に該当するのかのアノテーションを行うことが必要となる。本研究では、PDF に対して直接アノテーションを行い、結果をテキスト形式で取り出すことができる PDFAnno [19] を用いて、論文 PDF からアノテーション済みのテキストを抽出する[2)]。具体的には、論文 PDF 上の各ページごとに例文として抜き出すべき行の範囲のアノテーションを行った上で、本文とアノテーションを PDF に埋め込まれている単語ごとの座標情報と共にテキストに出力する。そして、その出力テキストを行単位に分割することで、情報抽出モデルの訓練・評価用データを用意する。

2）　PDFAnno のインターフェースについては、付録 A を参照されたい。

3.3 例文抽出

3.3.1 問題設定

化学分野等では、すでに固有表現抽出や関係抽出の技術を用いて論文から情報抽出を行う試みがある。本研究ではその情報抽出の技術を言語学分野に応用することで、言語学論文からの例文の自動抽出を試みる。

ここで、論文 PDF から例文を抽出するためには、前節で用意されたアノテーション済みデータについて、例文が含まれる行の範囲を予測する必要があるが、本研究では、固有表現抽出の手法として一般的な BIO-tagging の手法 [20, 21] の変種である BIEOS-tagging の手法を用いて例文が含まれる行のスパン予測を行う。具体的には、例文の先頭となる行に B (Beginning)、例文中となる行に I (Inside)、例文の範囲外には O (Outside)、例文の最終行に E (End)、さらに 1 行からなる例文の行に S (Single) というラベルを予測するモデルを学習することにより、例文が含まれる行のスパンを予測し例文を抽出する。

3.3.2 モデルアーキテクチャ

固有表現抽出を行うモデルとしては、ゲート付き畳み込みニューラルネットワーク (GCNN; Gated Convolutional Neural Network) [22] [3] と CRF を組み合わせたモデルを使用する。

本研究で用いる GCNN の隠れ状態 $\boldsymbol{h}_0, \cdots, \boldsymbol{h}_L$ は以下のように計算される。

$$\boldsymbol{h}_l(\boldsymbol{h}_{l-1}) = \boldsymbol{h}_{l-1} \otimes \sigma(\boldsymbol{h}_{l-1} * W + b)$$

ここで、$\boldsymbol{W}$, b は学習されるパラメータであり、$\boldsymbol{h}_0$ は、入力ベクトル系列 $\boldsymbol{X} = x_0, \cdots, x_N$ に相当する。また、σ はシグモイド関数であり、$\otimes$ はアダマール積を表す。

また CRF (linear-chain CRF) は、GCNN により得られた入力系列の特徴ベクトルを入力とし、最も尤度が高い BIEOS ラベルの系列を出力するモデルである。

$$P(\boldsymbol{y}|\boldsymbol{h}_t, \boldsymbol{W}, b) = \frac{\prod_{t=i}^{n} \exp(\boldsymbol{W}^{y_{t-1}, y_t} \boldsymbol{h}_t + b)}{\sum_{y'} \prod_{t=i}^{n} \exp(\boldsymbol{W}^{y_{t-1}, y_t} \boldsymbol{h}_t + b)}$$

ここで、$\boldsymbol{W}$, b は学習されるパラメータである。

3) ゲート関数は本論文とは異なり、$h_l(\boldsymbol{h}_{l-1}) = (\boldsymbol{h}_{l-1} * W + b) \otimes \sigma(\boldsymbol{h}_{l-1} * V + C)$ で隠れ状態が計算される。ただし、$\boldsymbol{W}, \boldsymbol{V}, b, c$ は学習されるパラメータである。

ラベルの予測の際には、1) 各行の単語列の埋め込み行列と単語の座標情報の埋め込み行列の連結を入力として、GCNN を用いて各行に対応する特徴ベクトル $\boldsymbol{h}^1_{row}, \cdots, \boldsymbol{h}^N_{row}$ を計算したのちに、2) さらにその各行の特徴ベクトルを入力として、GCNN を用いて周辺の行の情報を考慮した各行のより良い特徴ベクトル $\boldsymbol{h'}^1_{row}, \cdots, \boldsymbol{h'}^N_{row}$ を計算し、3) 最後に、そのベクトルを入力として CRF で BIEOS ラベル $y_1, \cdots, y_N$ の予測を行う。情報抽出モデル全体のアーキテクチャについては、付録 B を参照されたい。

また、モデルのアーキテクチャについては今後検証を進めていく段階で変更する可能性も考えられる。例えば、GCNN の部分を双方向 LSTM (Bidirectional Long Short-Term Memory) モデルに変更することなどが考えられるが、モデルの最終的な選定は今後の課題としたい。

3.3.3 予備実験

JEAL・LI・NLLT の各論文の 2 年間（2006 年-2007 年）で収録されている日本語統語論に関する論文すべて（各 5 報の合計 15 報、合計 381 ページ）を学習・訓練データとして、予備実験を行った。全 381 ページのうち 90% を学習データ、残りの 10% を評価データとした。すべてのデータについて、例文の先頭となる行に B (Beginning)、例文中となる行に I (Inside)、例文の範囲外には O (Outside)、例文の最終行に E (End)、さらに 1 行からなる例文の行に S (Single) のラベル付けがされている。学習は 10 エポック行った。

表 2 予備実験の結果

		Precision	Recall	F1-score
Tag	O	0.990	0.995	0.992
	B-example	0.929	0.844	0.884
	I-example	0.871	0.950	0.909
	E-example	0.942	0.844	0.890
	S-example	0.818	0.692	0.750
Span	example	0.800	0.750	0.774

タグレベルとスパンレベルでの結果を表 2 に示す。ここで、タグレベルでの評価では、ある行のモデルの予測が正解のラベルと一致していたら正解と見做されており、スパンレベルの評価では、ある例文の範囲全体に対するモデルの予測が正解のラベルと完全一致した場合にのみ正解と見做されている。

スパンレベルの評価を見ると、15 報という少量なデータにも関わらず 8 割程度の精度を達成できているという結果になった。また、タグレベルでの評価において S-example の精度が悪くなっているのは、その事例数が少ないためと考えられ、学習データを増やすことによって解決し得る問題であると考えられる。

3.4 後処理

前節までにより、対象の論文から例文が自動で抽出されるが、一般に言語学の論文の例文は以下のような形で提示される。

(1) Taroo-ga Hanako-ni/*o au.
Taroo-Nom Hanako-Dat/Acc see
'Taroo sees Hanako.' (Takahashi, 2006) [23]

各行はそれぞれ上から 1) 例文、2) グロス[4]、3) 英語訳に対応する。このように、各行はそれぞれ性質の異なるテキストとなっているため、抽出後に後処理的に分割が必要となる。また、英語で書かれた論文を対象とする際には、例文は漢字やひらがなではなくローマ字を用いて提示されているため、それらを漢字平仮名混じりの例文へと変換する必要がある。さらに、(1) の例文には-ni/*o という部分があるが、「に」を用いるのは文法的だが、「を」を用いるのは非文法的であること[5]、すなわち「太郎が花子に会う。」は正文であるが「太郎は花子を会う。」は非文であるということを、この 1 行で示している。こちらについても、データセットに含めるためには後処理的に 1 行から 2 文への変換を行う必要がある。

3.5 今後の展望

以上で紹介した手法を用いて、その規模に課題を抱えている JCoLA[10] の例文数を、まずは既存の言語学論文の例文に基づくデータセット [6, 8] に匹敵する 10,000 文規模に拡張する予定である。その後、必要に応じてさらにデータセットの規模を拡大していく予定である。

一方、言語学の論文には純粋な統語的制約の違反によって容認不可能とされている例に加えて、ある特定の解釈をすることができないという意味において容認不可能とされている例もある。

(2) John no hon-o kaita.
John-GEN book-ACC wrote
'John wrote a book.' Sudo (2015) [24]

ここで、(2) は「ジョン」が「書いた」という動作の主体であると解釈される場合にのみ容認不可能な例であるが、言語モデルにこの文の容認度を予測させる際に、ある特定の解釈を強制することは難しく、したがってデータセットに含めるのが必ずしも適切ではない例となる。以上のような例への対応については、今後の検討事項としたい。

4 おわりに

本論文では、すでに他分野で応用が進んでいる情報抽出の技術を言語学の分野に応用し、理論言語学の論文から自動で例文を抽出することで、言語モデルの統語的評価のためのデータセットを作成する試みを紹介した。既存の言語モデルの統語的評価用データセットは、例文を手動で抽出していたことによりデータの規模を容易に拡大できないという課題 [6, 8] や、例文を自動で生成していたことにより必ずしも複雑な統語現象を対象にした検証ができるデータセットではないという課題 [7, 9] を抱えていた。

本論文で紹介した言語学論文からの例文の自動抽出が実現すれば、理論言語学の論文で問題となっているような複雑な統語現象を扱い、かつ大規模であるという既存のデータセットにはなかった特徴を持つ言語モデルの統語的評価のためのデータセットを作成することが可能になるだろう。

謝辞

本研究は、JST さきがけ JPMJPR21C2 の支援を受けたものです。

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł Ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30, pp. 5998–6008. Curran Associates, Inc., 2017.

[2] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In **International Conference on Learning Representations**, 2019.

[3] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. **CoRR**, Vol. abs/1905.00537, ,

4) グロスとは、例文中の各形態素ごとに付与された逐次訳である。
5) 言語学ではある例文が非文法的（容認不可能）であることを*のマークで示すのが慣例である。

2019.

[4] Noam Chomsky. **Syntactic structures**. Mouton, 1957.

[5] Martin B H Everaert, Marinus A C Huybregts, Noam Chomsky, Robert C Berwick, and Johan J Bolhuis. Structures, not strings: Linguistics as part of the cognitive sciences. **Trends Cogn. Sci.**, Vol. 19, No. 12, pp. 729–743, December 2015.

[6] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. **Transactions of the Association for Computational Linguistics**, Vol. 7, pp. 625–641, November 2019.

[7] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. BLiMP: The benchmark of linguistic minimal pairs for english. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 377–392, December 2020.

[8] Daniela Trotta, Raffaele Guarasci, Elisa Leonardelli, and Sara Tonelli. Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus. In **Findings of the Association for Computational Linguistics: EMNLP 2021**, pp. 2929–2940, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[9] Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. CLiMP: A benchmark for Chinese language model evaluation. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 2784–2790, Online, April 2021. Association for Computational Linguistics.

[10] 染谷大河, 大関洋平. 日本語版 CoLA の構築. 言語処理学会第 28 回年次大会, 2022.

[11] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of LSTMs to learn Syntax-Sensitive dependencies. **Transactions of the Association for Computational Linguistics**, Vol. 4, pp. 521–535, December 2016.

[12] Alec Marantz. Generative linguistics within the cognitive neuroscience of language. Vol. 22, No. 2-4, pp. 429–445, 2005.

[13] Tal Linzen and Yohei Oseki. The reliability of acceptability judgments across languages. **Glossa: a journal of general linguistics**, Vol. 3, No. 1, 2018.

[14] Rebecca Marvin and Tal Linzen. Targeted syntactic evaluation of language models. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 1192–1202, Brussels, Belgium, 2018. Association for Computational Linguistics.

[15] Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. Cross-linguistic syntactic evaluation of word prediction models. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics.

[16] Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. Colorless green recurrent networks dream hierarchically. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, pp. 1195–1205, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[17] Shauli Ravfogel, Yoav Goldberg, and Francis Tyers. Can LSTM learn to capture agreement? the case of basque. In **Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 98–107, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[18] Aixiu An, Peng Qian, Ethan Wilcox, and Roger Levy. Representation of constituents in neural language models: Coordination phrase as a case study. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 2888–2899, Hong Kong, China, November 2019. Association for Computational Linguistics.

[19] Hiroyuki Shindo, Yohei Munesada, and Yuji Matsumoto. PDFAnno: a Web-based Linguistic Annotation Tool for PDF Documents. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Miyazaki, Japan, May 7-12, 2018 2018. European Language Resources Association (ELRA).

[20] Lance Ramshaw and Mitch Marcus. Text chunking using transformation-based learning. In **Third Workshop on Very Large Corpora**, 1995.

[21] Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In **Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)**, pp. 147–155, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[22] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In **Proceedings of the 34th International Conference on Machine Learning - Volume 70**, ICML'17, p. 933–941. JMLR.org, 2017.

[23] Daiko Takahashi. Apparent parasitic gaps and null arguments in japanese. **J. East Asian Ling.**, Vol. 15, No. 1, pp. 1–35, 2006.

[24] Yasutada Sudo. Hidden nominal structures in japanese clausal comparatives. **J. East Asian Ling.**, Vol. 24, No. 1, pp. 1–51, 2015.

[25] Hideki Kishimoto. Ditransitive idioms and argument structure. **J. East Asian Ling.**, Vol. 17, No. 2, pp. 141–179, 2008.

A　PDFAnno のユーザーインターフェース

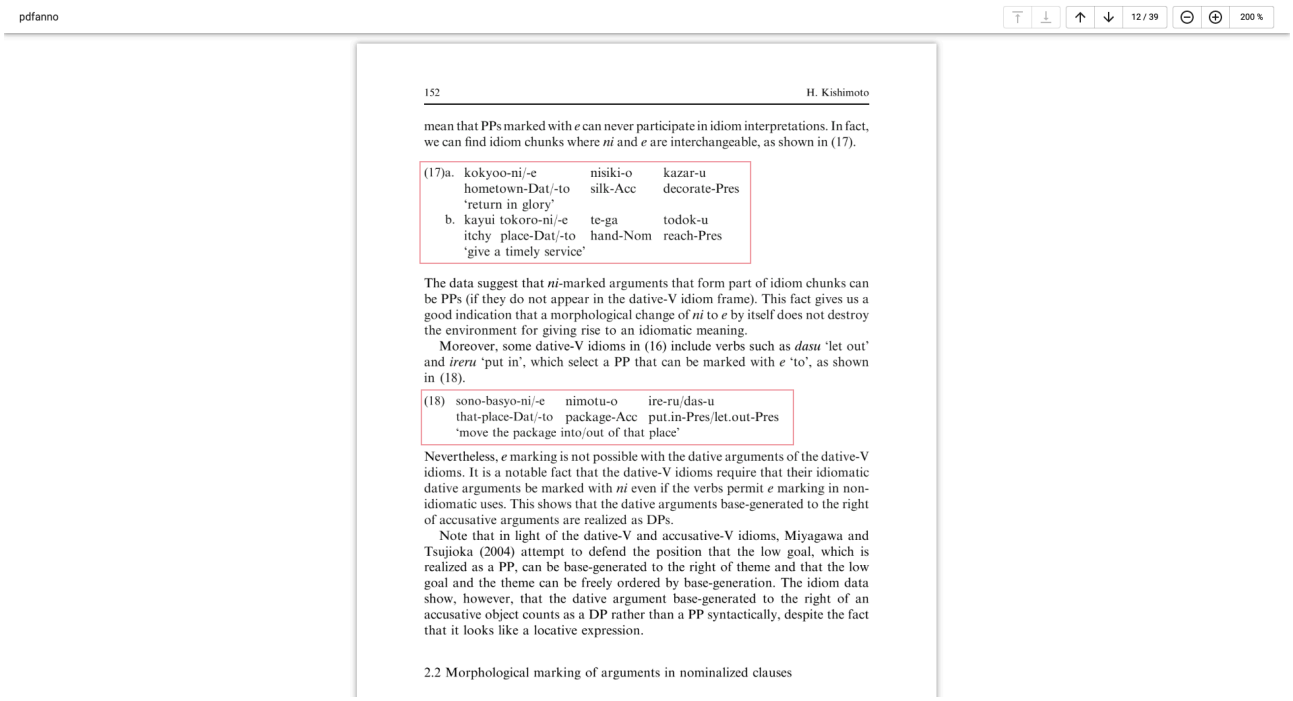

図 1　PDFAnno のインターフェース。例文の範囲を PDF に直接アノテーションできる。論文は Kishimoto (2008) [25]。

B　モデルアーキテクチャ

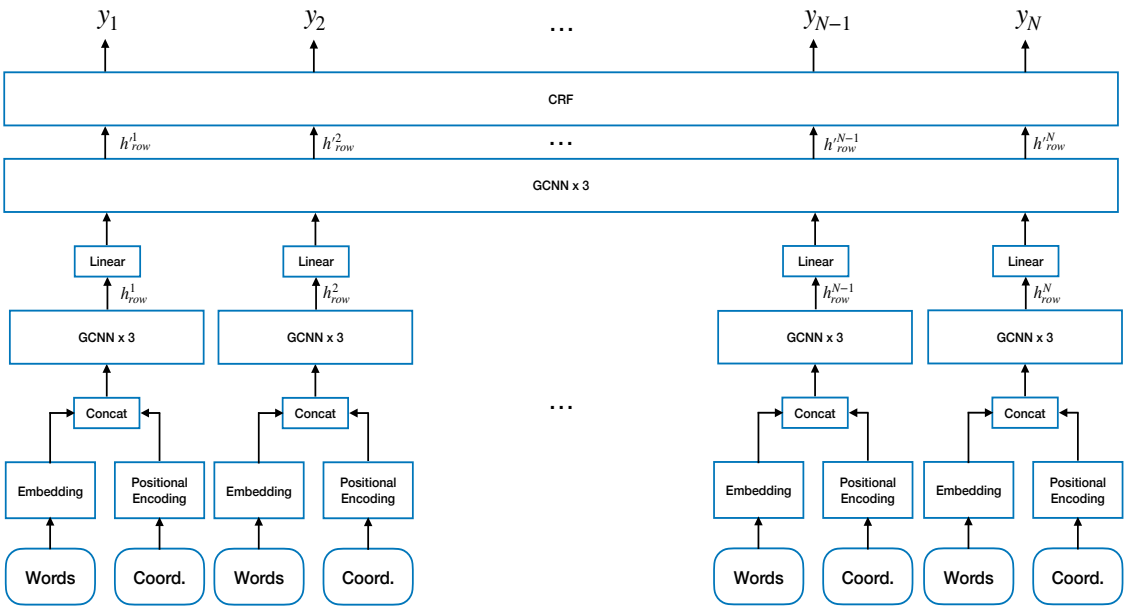

図 2　情報抽出モデル全体のアーキテクチャ。Coord. は PDF 文書に埋め込まれている単語の座標情報である。