名詞句の処理に頑健な構文解析器

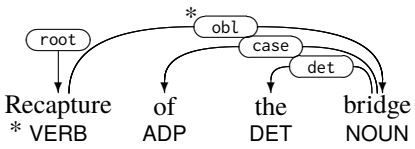
金山 博* 岩本 蘭 † 村岡 雅康* 大湖 卓也* 宮本 晃太郎*
* 日本アイ・ビー・エム株式会社 東京基礎研究所 † 慶應義塾大学 理工学研究科
* {hkana,mmuraoka,ohkot,kmiya}@jp.ibm.com † r.iwamoto@keio.jp

1 はじめに

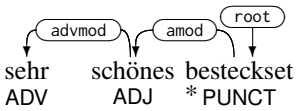
レビューの分析や契約文書の処理など応用の局面において、意味解析や情報抽出を目的とした実文書の正確な構文解析が求められている。近年は多言語の Universal Dependencies (UD)[4, 5] のコーパスが利用しやすくなっており、構文解析器（品詞タグ付け・依存構造解析）の訓練に用いられている。

しかし、実世界の文書に対する UD で訓練された解析器の出力を見ると、英語 (1)・ドイツ語 (2) [1] の例のような誤解析が目立つ[2]。

(1)



(2)



いずれも入力したテキストは短い名詞句であるが、(1) では命令法の文であるかのように 1 語目が動詞としてタグ付けされ、4 語目との間の関係ラベルも誤っている。(2) は名詞が大文字で書かれていないせいか、3 語目の品詞が句読点であると認識されてしまっている。これらは、UD の shared task[12] で高い精度を示している Stanza [8] の出力における例であるが、UDPipe[11] や Spacy[2] など、他の構文解析器においても同様の事象が見られた。

訓練データの多くが定動詞を含む文であって、名詞句となっているものが少ないとしたら、(1)(2) のような名詞句に対する品詞タグ付けや構文解析の誤りは多くなるであろう。また、(2) の異常にも見える誤解析は、訓練データのほとんどがピリオドで終わる文であるためにその傾向が強く学習されているためだと推測される。すなわち、これらの誤解析は、訓練データと解析対象のテキストの傾向の違いの問題と捉えることができる。

本稿では、これらの点、すなわち名詞句と文末のピリオドに注目し、UD のコーパスと実世界の入力における傾向の違いや解析時の影響を調べるとともに、新たに人手でアノテーションを加えることなく既存のコーパスを拡充して解析器を頑健化する方法を提案し、構文解析器やその後段のアプリケーションにおける変化を観察する。

2 定義

本稿で扱うデータや言語現象に関する用語について定義しておく。

処理単位 コーパス中で一文とみなされるテキスト[3]、または品詞タグ付けや依存構造解析の際に（しばしば文区切りの処理を施した後の）入力の単位となるもの。後述の「文」と区別するためにこのように呼ぶ。

文 定動詞を含むなど、句構造文法において非終端記号 S (Sentence) を構成するような処理単位またはその部分。本稿ではその厳密な議論はしないが、後述の「名詞句」は「文」ではない典型例である。

名詞句 処理単位の全体または部分において、その主辞の品詞タグが名詞（NOUN）または固有名詞（PROPN）であり、かつその主辞に be 動詞などのコピュラ（依存関係ラベルが cop であるもの）が係っていないもの[4]。

句点 処理単位の最後のピリオドを、簡単のため句点と呼ぶ[5]。

句点無 処理単位の最後の語が ‘.’・‘!’・‘?’ などの約物（やくもの）でない場合。すなわち、最後の語の品詞タグが PUNCT 以外である場合。

1） 逐語訳は ‘very nice cutlery’。
2） * は品詞タグまたは依存関係ラベルの誤りを示す。
3） UD で用いられる CoNLL-U フォーマットで “# text =” のメタデータとなる、通常は sentence と呼ばれる単位。
4） 内容語主導の UD の構文構造では、この考慮が必要となる。
5） 日本語等の「。」を含む。‘Mr.’ などのピリオドは含まない。

表 1　観察したデータと、名詞句・句点無の割合 (%)

言語	コーパス UD/review	名詞句 UD	名詞句 review	句点無 UD	句点無 review
en	EWT/SemEval	6.5	3.0	14.0	1.0
de	GSD/Gestalt	2.4	28.0	0.4	12.0
fr	GSD/SemEval	2.6	36.0	1.9	3.0
es	AnCora/SemEval	2.6	25.0	0.2	7.0

3　予備実験

表 1 は、UD とレビュー文のコーパスについて、処理単位が名詞句および句点無であるものの割合を 4 言語で調べたものである。UD は version 2.6 の訓練用セットのアノテーションから自動判定し、レビュー文はアスペクトベースの評判分析の shared task のデータ [7, 9] から各言語 100 例ずつを抽出して人手で検証した。

特にドイツ語 (de)・フランス語 (fr)・スペイン語 (es) の UD コーパスにおいて、名詞句や句点無の割合が非常に低く、コーパス作成時に処理単位として標準的な文が好んで選択されていることがわかる。英語 (en) の EWT コーパスは web 上のテキストが基となっているため、名詞句や句点無の割合が比較的高くなっている。レビューのデータでは名詞句や句点無の入力が多くなる傾向があるが、SemEval のデータセット（特に英語）はアスペクトの取得をする目的からも、ほとんどの文が句点で終わるなど、現実の入力よりは統制が取れたセットになっていた。

以上の結果より、言語やデータセットに依存するものの、UD とレビューのコーパスの間でも現象の分布が大きく異なることが確認できた。実世界のデータではさらに異なる傾向があると考えられる。

4　コーパスの拡張

本節では、3 節で確認したコーパスの分布の違いの問題を解決すべく、構文解析器の訓練コーパスを拡張する方法を示す。すなわち、訓練時における UD コーパスへのバイアスを減らして、実応用の際にも有用なモデルを作ることを目指す。

4.1　句点の除去

最も簡単な方法は、コーパス中の処理単位の末尾の句点を一定の割合で除去することである。これにより、句点の有無によって構文の構造が変わったり、通常の語に PUNCT の品詞タグが付与されたりする現象を防ぐことができると考えられる。

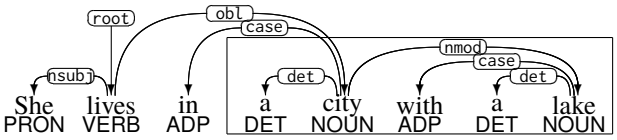


図 1　文から名詞句の部分木（枠内）を抽出する例

ほとんどの句点はそれに係る修飾語を持たないので、単純に句点を除去するだけで正しい構文構造を保てるが、UD_English-EWT のデータの中には例外的な事象[6]があるので、修飾語を持つ場合には除去の対象としないようにする。

4.2　名詞句の追加

定動詞を主辞とするような典型的な文に加えて、名詞句で完結する処理単位の割合を増やすために、元のコーパスから名詞句である部分木を抽出してコーパスに加える処理を行う。その手順は以下の通りである。

- 処理単位の中から、主辞でない部分にある名詞句を検出する。名詞句の同定の方法は 2 節の定義に従う。
- そのうち、4 語未満からなるものは構文構造の学習に適さないため除外する。
- UD の内容語重視の依存構造では、名詞句の部分木の中に前置詞 などの機能語が含まれうるため、名詞句の主辞に case, punct のラベルで係る語がある場合には、その部分を除外する。図 1 の例のように、‘city’ の部分木から ‘in’ を除いた部分を名詞句とする。
- こうして抽出した名詞句をプールして、元の文数に対して一定の割合でランダムに選択して訓練コーパスに追加する（元からのデータはそのまま保つ）。

5　実験

5.1　実験設定

拡張コーパスを使った訓練に基づく品詞タグ付けと依存構造解析の変化を観察するために、4 つの言語の名詞句を集めたデータを以下の手法により作成した。

6）UD_English-EWT の訓練データ中に、接続詞 ‘and’ が文末のピリオドに係る場合などがあった。

表 2　Wikipedia セクション名のデータの例

en	all passenger trains
	cobordism of manifolds with additional structure
fr	ponts sur d'autres cours d'eau
	instance vérité et dignité
de	Meine Daten und ich
	Mangelnde wissenschaftliche Grundlage
es	recopilatorios y discos especiales
	contenido de agua en el suelo

- Wikipedia の各言語版のセクション名[7]の文字列のうち、3 語以上からなるものを列挙
- 句点などの記号が含まれるものを除外
- 言語判定で他の言語と推定されたものを除外
- ドイツ語以外は先頭の文字を小文字化
- 先頭の 1 語が重複するものが 3 個以上は無いようにして、各言語 1500 個をランダムに取得

こうして得られる文字列のほとんどは名詞句であると推定できる。表 2 に 4 言語における例を示す。これらを解析して、その主辞（root）となる語が名詞となる割合（NP 検出率）と、最後の語の品詞が PUNCT 、または依存関係ラベルが punct となるような誤解析の数（誤句点数）を調べる。

さらに、構文解析の intrinsic な評価として、UD のテストコーパスでの依存構造解析の精度 (LAS) を測定する。訓練データとテストデータの分布に差が出ることにより、その値は理論的には低下するため、その減少幅が最小限であれば成功といって良い。

また、extrinsic な評価として、構文解析の結果を使った評価表現抽出 [3, 14] を行う。解析対象のデータは 3 節で観察した、SemEval 等の shared task のデータ [7, 9] の約 500 文である。評価表現の指標である適合率・再現率のうち、適合率は辞書の性質や否定等の扱い方の正しさに依存する傾向があり、再現率は辞書の充実の度合いのほか、主辞の同定の正確さに左右される。今回の実験では適合率を変化させる要因は少ないので、名詞句関連の構文構造の影響を受ける再現率を、評価指標として用いる。

5.2　結果と考察

ドイツ語・フランス語・スペイン語・英語の UD コーパスの訓練データに対して、4 節に示した 2 つの改変を施した拡張コーパスを用いて、Stanza [8] の再学習を行った。いずれの場合も、tokenize, mwe, lemmatize のモデルは標準のまま固定し、pos と depparse のモデルを max step=5000 の設定で学習させている。

コーパスを拡張する際に、m% の割合で文末の句点を除去、n% の名詞句を追加する。$m = 0, n = 0$ は元の UD コーパスをそのまま使って訓練する場合であり、各言語のベースラインとなる。m, n の値を変えて訓練したモデルで、Wikipedia のセクション名のデータにおける NP 検出率と誤句点数、UD のデータでの LAS と、評価表現抽出（SA）の再現率を測定する。Stanza の学習ではランダムの要素があり、試行ごとに精度が異なるため、$m = 0, n = 0$ の場合のみ 10 回の訓練を行って、それぞれのモデルで測定した際の平均値と標準偏差を示す。これらの結果が表 3 である。

4 言語すべてにおいて、$m = 0, n > 0$ のすべての場合で NP 検出率に改良が見られ、また n の値を増やすと NP 検出率がほぼ漸増する。このことから、名詞句の性質は異なっていても[8]、名詞句である処理単位の追加によって名詞句の解析結果を改善できることがわかる。$n = 100$ としても名詞句として解析されないものの中には、そもそも動詞句であるものも含まれていた。

誤句点数の減少を見ると、どの言語でも、少量の改変によって PUNCT に関連する誤りをほぼ防げるようになった。文末の句点を除去した場合 ($m > 0$) だけでなく、名詞句の追加 ($n > 0$) によっても誤句点数を減らせているのは、追加した名詞句に句点が基本的に含まれていないことによる効果である。

以上は Wikipedia のセクション名における観察の結果である。処理単位が名詞句である場合の解析結果を改善したことにより、一般的な文の解析の性能が低下することが懸念されるが、UD のテストデータの LAS の値を見ると、ベースラインより標準偏差を超えて下がっているケースは、m や n が大きすぎる場合に限られており、LAS がほぼ不変であったり、むしろ向上している場合も見られる。なお、$m = 100$ とした場合には、すべての文末の句点を除去して訓練をすることになるが、テストデータには文末の句点が存在するため、LAS が大きく下がるのは想定通りである。

評価の方法が難しいものの、最も重視したいの

7）記事のタイトルは固有名詞がほとんどなのでテストに不適であった。

8）UD コーパスの部分木と Wikipedia セクション名のデータは完全に独立である。

表 3 4 言語で、UD の訓練コーパスに対して文末の句点を m% 除去、名詞句を n% 追加して構文解析器を再訓練をした際の、各データ上での構文解析および評価表現抽出の結果（指標は誤句点数を除いて % 表記。）。 $m = 0, n = 0$（ベースライン）のみ 10 試行の平均値と標準偏差を示す。太字はベースラインに対して標準偏差を超える改善が見られた場合。

ドイツ語 (de)

		セクション名		UD	SA
m	n	NP 検出率	誤句点数	LAS	再現率
0	0	97.4	3.2	79.68	52.1
		±0.16	±1.75	±0.25	± 1.0
0	10	**98.1**	**0**	79.59	52.9
0	20	**97.7**	**0**	79.64	**53.8**
0	50	**98.1**	**0**	79.23	51.5
0	100	**98.4**	**0**	79.60	51.5
10	0	97.3	**0**	79.85	**53.2**
10	10	**97.8**	**0**	79.87	**54.6**
20	0	97.1	**0**	78.98	51.0
20	10	97.5	**0**	**80.20**	52.9
50	0	97.3	**0**	79.73	50.4
50	50	**97.9**	**0**	79.70	52.4
100	0	97.4	**0**	76.78	49.6

フランス語 (fr)

		セクション名		UD	SA
m	n	NP 検出率	誤句点数	LAS	再現率
0	0	91.4	4.2	87.14	43.0
		±0.55	±0.55	±0.18	± 0.5
0	10	**93.2**	3	**87.33**	42.0
0	20	**93.2**	**0**	87.25	**44.0**
0	50	**94.4**	**0**	86.76	42.6
0	100	**95.5**	**0**	86.37	**43.6**
10	0	**92.5**	**1**	87.01	42.6
10	10	**93.2**	**0**	87.09	43.0
20	0	90.6	**0**	87.31	43.2
20	10	**92.9**	**0**	87.19	42.4
50	0	91.1	**0**	**87.57**	**43.6**
50	50	**93.6**	**0**	87.02	42.8
100	0	**92.3**	**0**	84.57	42.0

スペイン語 (es)

		セクション名		UD	SA
m	n	NP 検出率	誤句点数	LAS	再現率
0	0	91.5	4.1	87.58	37.5
		±0.68	±2.90	±0.16	± 0.6
0	10	**93.1**	**1**	**88.21**	37.2
0	20	**92.8**	**1**	**88.02**	37.8
0	50	**94.2**	**1**	87.37	**38.4**
0	100	**94.7**	**0**	87.59	**38.2**
10	0	91.3	**0**	87.63	37.8
10	10	**92.7**	**0**	87.67	36.8
20	0	**93.5**	**0**	87.28	36.4
20	10	**93.1**	**0**	87.28	37.6
50	0	91.0	**0**	87.52	38.0
50	50	**94.4**	**0**	87.52	38.0
100	0	91.9	**0**	86.83	37.8

英語 (en)

		セクション名		UD	SA
m	n	NP 検出率	誤句点数	LAS	再現率
0	0	91.6	0.7	83.84	48.9
		±0.63	±0.67	±0.14	± 0.9
0	10	**93.9**	1	**84.09**	49.0
0	20	**94.6**	**0**	83.91	49.6
0	50	**95.3**	**0**	83.88	48.6
0	100	**95.3**	**0**	**84.00**	48.8
10	0	91.7	**0**	83.81	47.6
10	10	**94.2**	**0**	83.71	49.0
20	0	91.1	**0**	**84.06**	49.2
20	10	**93.7**	1	83.96	49.6
50	0	91.4	2	**84.03**	49.4
50	50	**95.4**	**0**	83.75	47.6
100	0	90.1	**0**	83.46	49.2

が、応用時における解析器の頑健性を測るための、評価表現抽出の再現率である。言語により最適となる m, n の値は異なっており、表 1 に示した各言語のコーパスにおける分布からその値を自動的に推定するには至っていないが、フランス語・ドイツ語・スペイン語においては m, n を 10〜20% 程度にすることにより、SA の再現率を上げられることがわかった。英語においては有意な差で再現率が向上したケースは無いが、これは 3 節の予備実験で見た通り、UD_English-EWT データ自体に適量のノイズが含まれていることと、SemEval のテストデータが逆に統制が取れた文が多いことから、名詞句の追加や句点の除去の効果が得づらいためだと考えられる。

6 まとめ

本研究では、名詞句と文末の句点に着目し、品詞タグ付けや依存構造解析における訓練コーパスの偏りを補正して、解析器を実応用に対して頑健化する実験を行った。句点の除去は非常に簡単な処理であり、句読点が構文解析に及ぼす影響は他の研究によっても指摘されている [10, 6, 13] が、名詞句を処理単位とした解析については、今後も追求する余地がある。今回は欧米言語のみを扱ったが、日本語などの言語においては、UD の構文を用いた名詞句の定義や抽出はより複雑であるため、さらなる調査や工夫が必要となろう。そして、UD コーパスに加えて生の文から訓練データを拡張する手法 [1] も合わせて、実用的な解析器の実現を進めていきたい。

参考文献

[1] Yousef El-Kurdi, Hiroshi Kanayama, Efsun Sarioglu Kayi, Vittorio Castelli, Todd Ward, and Radu Florian. Scalable cross-lingual treebank synthesis for improved production dependency parsers. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pp. 172–178, Online, December 2020. International Committee on Computational Linguistics.

[2] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020.

[3] Hiroshi Kanayama and Ran Iwamoto. How Universal are Universal Dependencies? Exploiting Syntax for Multilingual Clause-level Sentiment Detection. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pp. 4063–4073, 2020.

[4] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, 2016.

[5] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pp. 4034–4043, Marseille, France, May 2020. European Language Resources Association.

[6] Joakim Nivre and Chiao-Ting Fang. Universal Dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pp. 86–95, Gothenburg, Sweden, May 2017. Association for Computational Linguistics.

[7] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pp. 19–30, 2016.

[8] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 101–108, Online, July 2020. Association for Computational Linguistics.

[9] Josef Ruppenhofer, Roman Klinger, Julia Maria Struß, Jonathan Sonntag, and Michael Wiegand. IGGSA shared tasks on German sentiment analysis (GESTALT). In Gertrud Faaßand Josef Ruppenhofer, editors, *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pp. 164–173, Hildesheim, Germany, October 2014. Universität Heidelberg.

[10] Anders Søgaard, Miryam de Lhoneux, and Isabelle Augenstein. Nightmare at test time: How punctuation prevents parsers from generalizing. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 25–29, 2018.

[11] Milan Straka. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 197–207, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[12] Daniel Zeman, Filip Ginter, Jan Hajič, Joakim Nivre, Martin Popel, and Milan Straka. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Brussels, Belgium, 2018.

[13] 金山博. 読点に頼らない統計的構文解析. 情報処理学会 170 回自然言語処理研究会, 2005.

[14] 岩本蘭, 金山博. 多言語極性辞書の構築とその包括的評価. 言語処理学会第 26 回年次大会予稿集, March 2020.