実世界の文書に対する構文解析器の疑似評価

金山 博　　宮本 晃太郎
日本アイ・ビー・エム株式会社 東京基礎研究所
{hkana,kmiya}@jp.ibm.com

概要

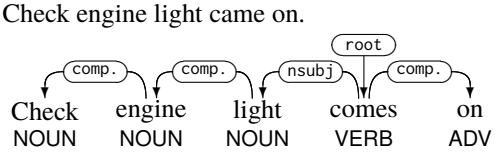
本研究では、構文解析器の実応用上での性能を、ドメイン特有の名詞句の構造に着目して推定する手法を示す。構文構造を付与したコーパスに比して、当該ドメインの専門用語のリストは低コストで取得でき、それらが部分木を形成するかをもって、解析器が重要な句を認識する能力を調べる。これによって、ドメインに特化した観点で複数の解析器の性能を比較することや、実用性を大きく損なう現象に注目して解析器をチューニングすることが可能となる。2 つのドメインで 5 つの解析器の性能を測る実験では、一般のベンチマーク上のスコアとは異なる傾向が観測され、実用の観点で考慮すべき点が示唆された。

1　はじめに

文区切り・単語区切り・品詞タグ付け・依存構造解析などを行う自然言語の解析器はさまざまな場面で利用されており、Universal Dependencies (UD) [10, 9] など各言語のリソースが整備され、学習や評価に使われるようになった。しかし、実応用の場面で入力される文書は、特殊な語彙や文体を含むことが多く、それらに対して既存の解析器が好適な結果を出力できているとは限らない。例えば、自動車の不具合の報告の中の文 (1)[1] は、(1a) の構造[2] を持ち、このように “check engine light” が複合名詞句として認識されれば、事物と動作の関係が抽出できて、現象の把握に役立つ。

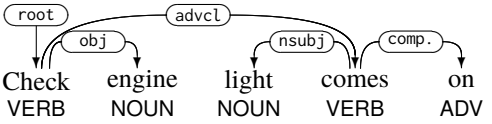
(1)　Check engine light came on.

(1a)



しかし、ある解析器の出力では (1b) のように、“check” が動詞とタグ付けされることにより名詞句の部分が部分木を構成せず、文全体の構造が乱れていた。これは後段の処理の大きな障害となりうる。

*(1b)



このように特定の分野の文書に対して解析器を適用して、意味役割付与 [1] や評価表現抽出 [6] などの後段の処理をする局面で、現存する解析器が充分な性能を出せているだろうか。その性能は UD のコーパス上でのベンチマークによって測ることができているだろうか。特に、複数の解析器やその様々なモデルのうち最も適したものを用いたい時に、どれを選択すれば望ましい出力が得やすいか。これらの質問に答えるための正確な測定には、当該分野の文書に構文構造を付与したコーパスを用いることが望ましいが、そのようなコーパスは存在しないし、それを逐一作成するコストは膨大である。

一方で、テキストマイニングの運用の局面では、しばしば当該分野で検索や分析の対象となる辞書を作成したり [7, 13]、既存のオントロジーを流用するなどして、名詞句や固有名詞が列挙されたリソースが存在することが多い。文 (1) における “check engine” ないし “check engine light” がそれに該当する。これらの句が当該分野の文書中に現れた場合、(1a) のように構文木の中で部分木を構成するはずである。一方、(1b) のように句が分断されている場合、解析器の明らかな出力誤りである場合が多い。

本論文では、分野特有の名詞句に対する解析結果の構造を調べることによって、構文構造のアノテーションに頼らずに解析器の実応用における性能を推定する方法を提案する。これによって、一般のベンチマーキングとは異なり、実応用の観点で複数の解析器の性能を比較することや、実用性を大きく損なう現象に絞って解析器のチューニングを行うことができるようになる。

1）「エンジンを確認せよという警告が点灯する」の意。
2）comp. は compound ラベルを示す。

表 1　分野名詞句の例。

自動車分野	契約書分野
warning light	private placement shares
check engine light	control termination
replacement tire	administrative agent
repair shop	restricted subsidiary
dealer shop	performance share units
Chrysler 200	equity interests

2　分野名詞句辞書

いわゆる複数語表現（MWE）で名詞的なもの [3] のうち、特定の分野で頻出する専門用語や固有名詞をここでは**分野名詞句**と呼ぶ。表 1 に、自動車の不具合のレポートと、サービスの契約書の各ドメインの分野名詞句の例を示す。

分野名詞句は次のような性質を持つ。

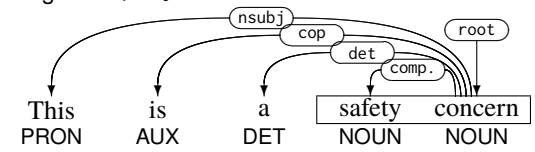
1. 専門用語や固有名詞のリストであるため、これらの語が文書中の表層上に現れた場合は、曖昧性が少なく、文脈により異なる解釈がされることは稀である。
2. これらの語は、情報検索や関係抽出など、当該分野の文書を処理する上で重要であり、品詞タグ付けや構文解析に失敗すると、構文解析結果の有用性が大きく損なわれる。

それらを列挙した辞書のことを、**分野名詞句辞書**と呼ぶ。分野名詞句辞書は、各分野の専門家が分析等の目的で既に作成済であることが多く、そうでなくてもテキストマイニングにより頻出する語やフレーズを集計するなどで半自動的に構築したり、重要キーフレーズの教師なし学習 [4] により獲得することができるなど、その構築自体に大きなコストはかからないという状況を想定している。

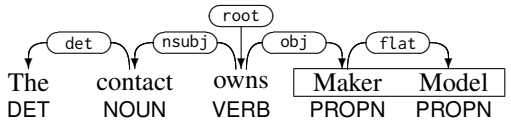
3　名詞句の構文構造

2 節で示した分野名詞句が、特定の分野の文の構文解析結果の中でどのような構造になっていると望ましいか、逆にどのような構造が出力されたら解析に失敗していると推定されるかを考える。そのために、分野名詞句と表層が一致した部分の構文木の形を以下の 6 種類に分類する。

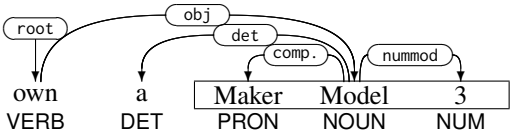
right　分野名詞句の多くは、例 (2) の “safety concern” の部分[3)]のように、compound のラベルで結ばれる複合名詞となっている。このように、分野名詞句の中で、最も右の語が主辞であり、その中の他のすべての語が名詞句内の語に係っている形を right と呼ぶ。

(2)


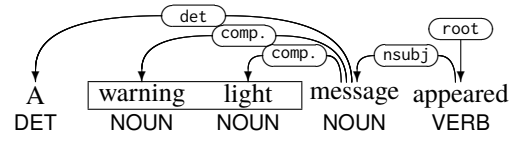
left　固有名詞の場合、(3) のように、flat や nummod のラベルで左側が主辞となる構造となることが一般的である。このように、最も左の語が主辞であり、その他の語が名詞句内の語に係る形を left と呼ぶ。

(3)


mid　3 語以上からなる分野名詞句の場合、それらの組み合わせにより、(4) のように主辞が中間に位置することがある。この形を mid と呼ぶ。

(4)


same　分野辞書にある名詞句が、さらに大きな名詞句の一部として出現する場合がある。構文解析器の出力 (5) では、“warning light” が分野名詞句として検出された。しかし “warning light message” がより大きな複合名詞であり、解析結果でも “warning” と “light” の双方が “message” に compound ラベルで係る形となっていることから、“warning light” に着目すると、その中で係り受けが閉じていない。厳密には “warning” と “message” の係り受けは正しくないが、名詞句の内部構造を正確に捉えることは難しく[4)]、“warning light message” が部分木を成せば全体の構文構造に破綻をきたすほどではないと考えて、エラーとしては検出しないこととする。このように、分野名詞句のそれぞれの語が外側の同じ語に係っている構造を same と呼ぶ。

(5)


3）　分野名詞句辞書に含まれる複合語と符合した部分を四角の枠で示す。

4）　UD のコーパスの中でも、3 語以上の flat や compound の内部構造は詳細にアノテートされていない。

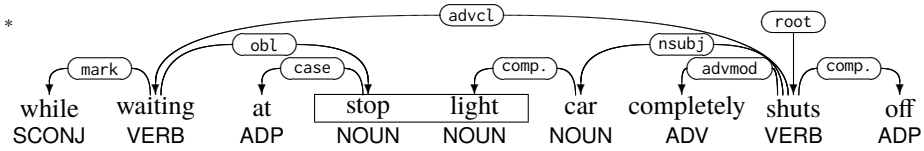
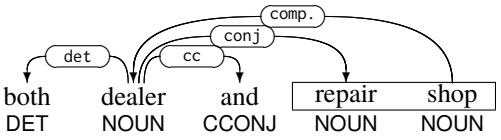


図 1　other となる例。分野名詞句 “stop light” の構造が分断されている。

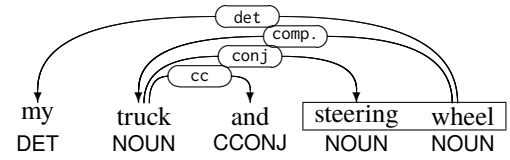
conj　並列句において、名詞句の構造が分断される場合も考慮する。(6) では、“repair shop” が分野名詞句であるが、“dealer” と “repair” が並列構造をなして “shop” に係る構造となっており、左側主辞の原則により係り受けが分断されているが、これは構文構造として正しい。このような場合、“dealer shop” が分野名詞句辞書にあるなら、並列構造が正しいとみなして、解析誤りではないと判定する。このようなケースを conj と呼ぶ。



(6)

other　これまでの分類に属さないものが other である。(1b) のように品詞タグ付けの誤りに起因するものや、図 1 のように名詞句が分断される場合などが該当し、これらを潜在的に重大な解析誤りであると捉える。

また、(7) の並列構造では、検出された分野名詞句 “steering wheel” の構文構造が分断されているが、これは “truck” と “steering” が並列とみなされるという誤った解析結果によるものである。“truck wheel” という語が分野名詞句辞書に無いことから、これも other に分類され、解析誤りであるとみなせる。



*(7)

4　評価実験

本節では、英語の 2 つの分野を題材に、3 節の基準で other と分類される分野名詞句の潜在的な解析誤りの量を調べることによって、当該分野における解析器の有用性を評価する。また、一般のベンチマークと比較して、各解析器の評価結果の傾向にどのような差異があるかを調べる。

表 2　解析器 A〜E の単語区切り (Words)・品詞タグ付け (UPOS)・係り受け解析 (UAS, LAS) の精度 (F1 値) を UD_English-EWT コーパスで測定したもの。

解析器	Words	UPOS	UAS	LAS
A	98.29	92.93	84.77	81.75
B	99.35	96.07	91.68	89.21
C	98.34	94.82	81.80	79.40
D	99.00	96.39	90.47	88.32
E	98.36	95.21	90.91	88.40

4.1　分野のデータ

まず、自動車分野のデータとして、National Highway Traffic Safety Administration (NHTSA) [2] が収集・公開している「Consumer Complaints」[5] のうちの 10,000 レコードを用いた。

また、法的契約文書のデータとして、CUAD [5] データセット[6] のうちの「2020」ディレクトリから抽出された 10,000 レコードを用いた。

分野名詞句辞書として、自動車分野では分析に用いていた既存の辞書にある 3,942 個の名詞句を活用した。契約書分野では文書集合中で連続して単語の先頭が大文字で書かれる頻度が高い名詞句 924 個を抽出した。それらの例は表 1 に示した通りである。

4.2　解析器

英語の解析器として、ここでは、UDPipe 2 [12] の UD_EWT 2.6 モデル、Stanza 1.0.0 [11]、Trankit 1.0.0 [8]、および 2 つの内製の解析器を含めた 5 つを比較する。なお、各解析器の優劣を議論することは本論文の趣旨ではないので、以下ではこれらの順序を変えて解析器 A〜E と表記する。

これらの解析器を UD_English-EWT のテストデータで評価した結果を表 2 に示す。品詞タグ付け・依存構造解析とも、B, D, E の値が高い。UAS, LAS が最も高い B は単語区切りの性能の高さによることもあり、D, E と本質的な性能に大きな差はない。

EWT コーパスにはさまざまな特殊な現象がある

5）https://www.nhtsa.gov/nhtsa-datasets-and-apis#complaints
6）https://paperswithcode.com/dataset/cuad

表 3　各解析器の自動車分野での名詞句の構造の頻度。括弧は other となる割合を示す。

解析器	right	left	mid	same	conj	other	
A	8405	475	99	185	2	201	(2.15%)
B	9131	101	71	218	2	89	(0.93%)
C	8865	122	88	398	0	114	(1.19%)
D	8954	99	55	242	3	137	(1.44%)
E	9083	96	40	253	2	83	(0.86%)

表 4　契約書分野での名詞句の構造の分類。

解析器	right	left	mid	same	conj	other	
A	5877	533	84	79	5	233	(3.42%)
B	6279	128	15	153	15	121	(3.11%)
C	6268	142	26	167	25	191	(2.80%)
D	6193	152	27	134	27	250	(3.68%)
E	6210	137	16	131	15	236	(3.50%)

ことが知られており [14]、UD のベンチマークと実応用での有用性には乖離がある可能性がある。この後の評価では分野のコーパスを用いて検証を行う。

4.3　評価結果

表 3 と表 4 に、2 つの分野の文書での分野名詞句の構造の頻度を示す。other となる場合が潜在的な解析誤りであるとして、その割合を付記してある。

解析器 A を除いて両分野で right が 90% を超えている通り、分野名詞句の大半は複合名詞として認識されている。A では他よりも left・mid が多いが、複合名詞が左主辞の固有名詞と解釈される場合がほとんどで、それらは分析に大きな支障はない。解析器 C は same が多く、より広い範囲の複合名詞と捉える傾向がある。

other の割合は契約書のほうが多い。フォーマルな文体ではあるものの長い文が多いことから解析が難しいことがわかる。また、conj の頻度に並列句の多さが現れている。

これらを表 2 で見た性能と比較すると、解析器 C は LAS のスコアが最も低いのに other の割合は小さく、特に契約書分野ではエラーの割合が最も低い。契約書分野では UD 上の係り受けの誤り率[7] と other の割合が負の相関を持っており、解析器の実データ上での性能は UD-EWT でのベンチマークでは測定できていないといえる。

7）1 から LAS の F1 値を減じたもの。

表 5　各構造と実際の誤りの関係。

	right	left	mid	same	conj	other
割合 (%)	93.3	1.9	0.7	2.7	0.0	1.3
正	40	35	38	39	33	2
誤	0	5	2	1	7	38

表 6　誤りの頻度の多い分野名詞句の例。

自動車分野	契約書分野
warning light	permitted other indebtedness
check engine light	consolidated EBITDA ratio
power steering	financial accounting standards
	codification topic

4.4　擬似評価の妥当性

other の検出による性能の評価が妥当であるかを知るために、各カテゴリの名詞句の構文構造が実際に正しいかどうかを、自動車分野の文に対する解析器 A〜E の解析結果から 40 件ずつサンプリングして調査した。表 5 にその結果を示す。これより、other の分類によって分野名詞句の誤りを検出することの適合率は 95% (= 2/40)、再現率は各カテゴリの頻度を考慮して 78.4% (= (49.8/40)/(63.5/40)) と計算され、簡単な方法でありながら名詞句周辺の解析誤りを正しく捉えられているといえる。

4.5　誤りの典型例

表 6 に、解析器 A が各分野で other となった分野名詞句のうち頻度が高いものを示す[8]。自動車分野では “warning” や “steering” などが動詞とみなされることによる誤りが目立つことなどがわかり、複合語を集約して一語とみなすなどの前処理によって解析誤りを防ぐことができる。契約書分野でも、解析に支障をきたす専門用語が見出された点が興味深い。

5　まとめ

本論文では、実応用上での解析器の性能を、分野特有の名詞句の解析結果に着目して推定する手法を示した。これにより、新たなコーパスを作成するコストをかけずに、一般のベンチマークとは別の観点で、当該分野の文書の解析において最も好ましい解析器やモデルを選択することができるようになった。また、後段の処理で致命的になりうる典型的な誤りを検知することができて、有用な解析器を実現するための示唆が得られた。

8）固有名詞は除外している。

参考文献

[1] Xavier Carreras and Lluís Màrquez. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the ninth conference on computational natural language learning (CoNLL-2005)*, pp. 152–164, 2005.

[2] Monica G. Eboli, Catherine M. Maberry, Ian A Gibbs, and Ramsi Haddad. Detecting potential vehicle concerns using natural language processing applied to automotive big data. In *Proceedings of the 26th International Technical Conference on the Enhanced Safety of Vehicles (ESV)*, 2019.

[3] Meghdad Farahmand, Aaron Smith, and Joakim Nivre. A multiword expression data set: Annotating non-compositionality and conventionalization for English noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*, pp. 29–33, Denver, Colorado, June 2015. Association for Computational Linguistics.

[4] Xiaotao Gu, Zihan Wang, Zhenyu Bi, Yu Meng, Liyuan Liu, Jiawei Han, and Jingbo Shang. UCPhrase: Unsupervised context-aware quality phrase tagging. In *KDD'21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 14-18, 2021*, Vol. 2021, 2021.

[5] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. CUAD: An expert-annotated NLP dataset for legal contract review. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, Vol. 1, 2021.

[6] 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.

[7] Tetsuya Nasukawa and Tohru Nagano. Text analysis and knowledge mining system. *IBM Systems Journal*, Vol. 40, No. 4, pp. 967–984, 2001.

[8] Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 80–90, Online, April 2021. Association for Computational Linguistics.

[9] 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.

[10] Joakim Nivre and et al. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of LREC*, Portorož, Slovenia, 2016.

[11] 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.

[12] Milan Straka and Jana Straková. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 88–99, Vancouver, Canada, August 2017.

[13] 那須川哲哉, 吉田一星, 宅間大介, 鈴木祥子, 村岡雅康, 小比田涼介. テキストマイニングのための辞書構築, 第 2 章, pp. 27–58. テキストマイニングの基礎技術と応用. 岩波書店, 2020.

[14] 金山博, 大湖卓也. UD_English-EWT との付き合い方. 言語処理学会第 28 回年次大会予稿集, March 2022.