複数の分野のコーパスを用いた述語項構造解析の比較

―『現代日本語書き言葉均衡コーパス』を用いて ―

吉本 暁文 （奈良先端科学技術大学院大学）[†1]
小町　守 （奈良先端科学技術大学院大学）[†2]
松本 裕治 （奈良先端科学技術大学院大学）[†3]

A Comparison of Predicate Argument Structure Analysis on Multi-domain Corpora

— Using the Balanced Corpus of Contemporary Written Japanese —

Akifumi Yoshimoto (Nara Institute of Science and Technology)
Mamoru Komachi (Nara Institute of Science and Technology)
Yuji Matsumoto (Nara Institute of Science and Technology)

1　はじめに

述語項構造解析とは、動作や状態、出来事を表す動詞、形容詞、動作名詞等を述語とし、その述語に関わっている単語や句とその役割を同定するタスクである。例えば「太郎が次郎に本を貸した」という文では「貸す」という単語が述語に相当する。また、「太郎」は本を貸すという動作の主体であり、「次郎」は動作の受け手であり、「本」は動作の対象であり、それぞれガ格、ニ格、ヲ格の項と呼ぶ。述語とこれらの項との関係を述語項構造という。述語項構造解析により、機械翻訳での単語の対応における誤りを低減したり、情報検索における絞り込みに役立てたりすることができるようになる。

しかしながら、従来の述語項構造解析は主に新聞記事を対象にして研究が進められてきた。新聞記事は日本語の書き言葉の中でも文体に揺れが少ないため、新聞記事を訓練・評価両方に用いた場合、高い性能を得られたとしても、それが他の分野においても同様に高い性能を示すとは限らない。

そこで本研究では、『現代日本語書き言葉均衡コーパス』（以下、BCCWJ と略す）を用いて新聞記事以外の分野を対象に述語項構造解析解析を行った結果について分析する。先行研究における述語項構造解析では、ほとんどの場合新聞記事を対象に評価されているが、本稿ではBCCWJ に含まれる Yahoo!知恵袋における評価結果について報告する。本研究では新たにタグ付けされた Yahoo!知恵袋[†4] を用いて述語項構造解析器を学習し、分野の異なるテキストに対して述語項構造解析を行なった結果について比較する。

[†1] akifumi-y@is.naist.jp
[†2] komachi@is.naist.jp
[†3] matsu@is.naist.jp
[†4] http://chiebukuro.yahoo.co.jp/

本研究の主要な貢献は、以下の 2 点である。

- Yahoo!知恵袋の述語項構造解析アノテーションを用いて述語項構造解析器のトレーニング・テストを行った初めての研究である。
- コーパスのドメインごとに傾向の異なるトレーニングデータが得られることを確認した。

本稿は以下のように構成される。まず 2 節で述語項構造解析で用いられているコーパスについて概観し、3 節で本研究で用いた述語項構造解析器について説明する。そして 4 節で評価実験の内容と定量評価の基準について述べ、5 節で結果の報告と分析を行う。最後に、6 節でまとめと今後の方針について述べる。

2 日本語述語項構造タグ付きコーパス

これまでに様々なジャンルのテキストを対象に、日本語の述語項構造タグ付きコーパスが開発されてきた。

たとえば、京都大学テキストコーパス Version 4.0 [Kaw02] は 1995 年 1 月 1 日から 17 日までの全記事（約 2 万文）と 1 月から 12 月までの社説記事（約 2 万文）の計 4 万文のうち 5,000 文に対して格・省略・照応・共参照の情報がアノテートされている。述語項構造情報は必須格・任意格の両方を表層格でタグ付けされている。また、KNB コーパス (Kyoto-University and NTT Blog Corpus) [橋本 11] は、「京都観光」、「携帯電話」、「スポーツ」、「グルメ」の 4 つのテーマについて書かれた 249 のブログ記事、4,186 文からなる人手による解析済みブログコーパスで、格・省略・照応情報がアノテートされている。アノテーション基準は京都大学テキストコーパスと共通である。京都大学テキストコーパス Version 4.0 と KNB コーパスを比較することで、同じアノテーション基準でアノテートされた新聞記事とブログ記事の述語項構造解析結果を比較することができるが、それぞれ 4,000–5,000 文とテストには十分な規模があるものの、機械学習手法で解析器を学習するには比較的小規模である。

一方、NAIST テキストコーパス 1.4 [飯田 10] は京都大学テキストコーパス 3.0 を元に、4 万文全体に対して照応・共参照・述語項構造の情報がアノテートされている。NAIST テキストコーパスでは述語の基本形に対し、格交替を原型に戻したうえで、必須格となる表現をガ格・ヲ格・ニ格でタグ付けしている。NAIST テキストコーパスは大規模に述語項構造がタグ付けされているものの、新聞記事という 1 つのジャンルでしかタグ付けされていないため、分野をまたいだ比較が難しい。

そこで、本研究では現代日本語書き言葉均衡コーパス (BCCWJ) を対象に、複数の分野における述語項構造解析器を比較した。BCCWJ にはコアデータ [小椋 09] と呼ばれる手作業で形態素情報・構文情報を付与したデータセットがあり、コアデータのうち、書籍 (PB)、新聞 (PN)、白書 (OW)、Yahoo!知恵袋 (OC) には [小町 11] により NAIST テキストコーパスと同じアノテーション基準による述語項構造と照応関係のタグ付けがなされている。

3　機械学習による述語項構造解析

これまでにいくつかの述語項構造解析手法が提案されている。また、述語項構造解析に関連するタスクとして述語の深層格と、述語がとる格を同定する格解析がある。たとえば、KNP [†1]は依存構造解析の過程で、大規模格フレームを用いた確率モデルによって格解析を行っている。他にも、[飯田 04, Ima09, Tai08, Yos11, 笹野 11] などがある。機械学習による手法は表層の語彙素性や大規模な格フレーム情報といった知識を用いることによって性能を向上させることができるが、一般的に機械学習による手法は解析器のトレーニングに用いたデータとテストに用いるデータが異なる分野であるとき、性能が落ちることが知られている。

そこで、本稿では、分野の異なるテキストに対する述語項構造解析器の性能を測るため、機械学習に基づく手法でトレーニングとテストにおいて異なる分野のテキストを用意し、それぞれで性能を比較する。具体的には、トーナメントモデル [飯田 04] によって項の同定を行う。トーナメントモデルでは、トレーニングの際には項（正解）である句と項でない句のペアに対し、その句のペアにおける手がかりから項である句を選ぶように機械学習を行う。テストの際には句のペアを作り、どちらが項候補かを分類してその項候補と残りの句のペアを作り、どちらが項候補かを分類するといった計算を繰り返す。これによって文全体から最終的な項候補を選出する。

本稿では分野の異なるテキストに対する項同定性能を比較するため、述語は既知、かつ文内に項が存在する事例のみをトレーニングとテストに用いた。[†2]

4　異なる分野のテキストを用いた述語項構造解析の比較実験

2 つの異なる分野における述語項構造タグつきコーパスを使用し、機械学習に基づく述語項構造解析器を用いて文内のガ格の項同定性能の比較実験を行なった。まず、それぞれの分野のコーパスでトレーニングした解析器をそれぞれの分野のコーパスでテストすることによって、分野による解析器の性能を評価する。次に、異なる分野のコーパスを追加して再学習することで、分野の異なるコーパスがテストにどのような影響を与えるのか調査する。

4.1　データ

BCCWJ 述語項構造・照応アノテーション[†3] に含まれる BCCWJ のコアデータに対する述語項構造アノテーションのうち、新聞記事 (PN)、Yahoo!知恵袋 (OC) のコーパスを用いて実験を行う。新聞記事データは 2012 年 4 月 4 日版のスナップショット、Yahoo!知恵袋データは BCCWJ 述語項構造・照応アノテーション v0.2（2012 年 10 月 5 日版）を用いた。

新聞記事データは 8,998 文、Yahoo!知恵袋データは 6,321 文である。

[†1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

[†2] 述語と句の間の係り受け関係の有無の判定には、人手による係り受け情報ではなく、自動解析結果を用いた。

[†3] http://cl.naist.jp/nldata/bccwj/pas/

表 1　文内ガ格の項同定実験に用いた素性

素性の種類	素性	詳細
文法的	品詞	“名詞-固有名詞”, “名詞-サ変接続” のような項候補の品詞
	格助詞	“は”, “が”, “を” のような項候補に続く助詞
意味的	固有表現	“ORGANIZATION”（組織名）などの固有表現の種類
	生物	項候補が生物である場合に 1
位置的	距離	項候補と述語との間の距離（... , -2, -1, 1, 2, ...）
	文頭	項候補が文頭にある場合に 1
	係り受け	項候補が述語と係り関係にある場合に 1
	連体	項候補が連体節の中にある場合に 1

4.2　ツール

形態素情報は BCCWJ コアデータに含まれる人手修正済みのデータを用いた。また、形態素解析済みのデータを入力として CaboCha 0.65[†4] UniDic モデルを用いて自動的に文節区切り・係り受け解析・主辞解析・固有表現解析を行った。分類器には LIBLINEAR 1.92[†5]を用いた。

4.3　素性

使用した素性は [飯田 04] の素性を基に一部改編したもので、表 1 に示す。これは元の素性から EDR 概念辞書や日本語語彙大系を用いる素性と項と述語の共起、項候補と照応関係にある名詞句の数と Salience Reference List を用いた素性を省いたものである。

4.4　評価尺度

項同定精度の評価には、適合率 (P)・再現率 (R)・F 値を用いた。$P = \frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fp}}, R = \frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fn}}, F = \frac{2 \times P \times R}{P+R}$ である。

5　結果と考察

5.1　素性の学習結果

各分野によって学習された素性の一部とその重みの絶対値を表 2 に示す。これらはトーナメントモデルにおいて句のペアから項候補を選ぶ際の選ばれやすさに影響する素性の一部である。素性は特に影響しているものとして句の主辞の品詞と名詞の分類と、直接の係り受け関係を示している。

[†4] https://code.google.com/p/cabocha/
[†5] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

表 2　項候補の選択に影響する素性と重み

OC				PN			
文頭側		文末側		文頭側		文末側	
代名詞	0.883	代名詞	0.395	代名詞	0.780	代名詞	0.350
名詞	0.154	名詞	0.082	名詞	0.476	名詞	0.166
固有名詞	0.556	固有名詞	0.325	固有名詞	0.493	固有名詞	0.481
普通名詞	0.345	普通名詞	0.209	普通名詞	0.203	普通名詞	0.179
直接係受	0.445	直接係受	0.665	直接係受	0.470	直接係受	0.709

5.2　項同定精度

分野の異なるテキストで文内ガ格の項同定の訓練・評価の違いを見るための比較実験の結果を、述語と項が係り受け関係があるかどうかで分け、表 3、表 4 に示す。

述語と項が係り受け関係にある場合、最も F 値が高くなるのは PN で訓練、PN で評価を行った場合となった。また、OC で訓練、PN で評価の場合が OC で訓練、OC で評価の場合よりも高い。PN は比較的 1 文あたりの項候補の数が多いと考えられるが、これらの結果から今回の実験設定では一番よく解析できる分野であることがわかる。一方で最も F 値が低くなったのは PN で訓練、OC で評価を行った場合となった。

PN で評価した場合について訓練時の各分野を比較しても訓練と評価を合わせた方が良い結果が得られることがわかり、OC で評価した場合についても同様である。

PN で訓練した場合について評価時の各分野を比較すると訓練と評価を合わせた方が良い結果が得られることがわかる。一方で、OC で訓練した場合については係り受け関係にある場合に PN による評価が良い結果を出しており、PN がこの場合よく解析できる分野であることがわかる。

また、OC の訓練データに PN の訓練データが加わると、OC の分野ではほぼ精度変化が見られなかった。一方で PN の訓練データに OC の訓練データが加わると、PN の分野では精度が低下した。PN の分野においては OC の訓練データがノイズになっている可能性が考えられるが、どのような特徴がノイズになっているのかの検討は今後の課題である。

述語と項が係り受け関係にない場合、OC で訓練した場合について評価時の各分野を比べると、係り受け関係ありの場合と異なり PN よりも OC で評価した場合の方が F 値は高くなり、係り受け関係がない場合は PN もあまり解きやすい分野ではなくなることがわかる。また、OC の訓練データに PN の訓練データが加わると、OC の分野では係り関係ありの場合と異なり精度の向上が見られた。今回の実験では項と述語の共起や意味カテゴリ、表層に関する情報を用いていないため、係り受け関係にない場合に項同定を行うための手がかりが不足している可能性が考えられるが、これらの素性を実装することによって、どのように傾向が変わるか調べるのは今後の課題である。

表 3　文内ガ格の項同定精度（述語と項が係り受け関係にある場合）

		評価					
		Yahoo!知恵袋 (OC)			新聞記事 (PN)		
		P	R	F	P	R	F
訓練	OC	0.690	0.901	0.781	0.722	0.877	0.792
	PN	0.665	0.908	0.767	0.777	0.917	0.841
	OC+PN	0.690	0.900	0.781	0.695	0.875	0.774

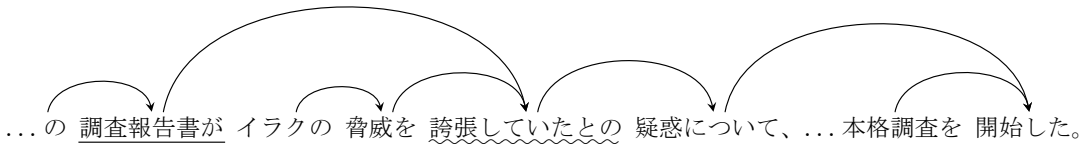
表 4　文内ガ格の項同定精度（述語と項が係り受け関係にない場合）

		評価					
		Yahoo!知恵袋 (OC)			新聞記事 (PN)		
		P	R	F	P	R	F
訓練	OC	0.462	0.533	0.495	0.388	0.596	0.470
	PN	0.426	0.449	0.437	0.524	0.698	0.598
	OC+PN	0.464	0.586	0.518	0.410	0.578	0.480

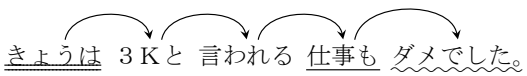
5.3　事例分析

以下に解析結果を示す。下線が正解の項、二重下線が不正解の句、波線が述語を表す。

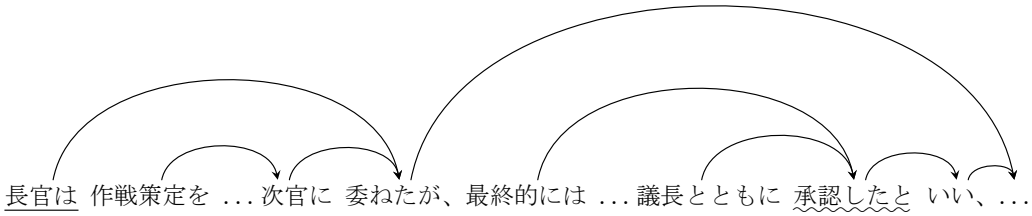
■述語と項が係り受け関係にある場合　これは訓練と評価にともに新聞記事を用いた事例である。PN 分野では物体は組織よりも動作主になりにくいと考えられるが、この事例では述語と項が係り受けの関係にあるため、ガ格の表層格である近くの項を捉えられている。



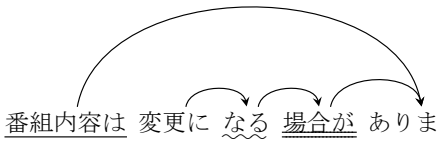
次の事例は、PN で訓練し、OC で評価した事例である。この場合、述語と項の間に係り受け関係があるにも関わらず、正解を出力できなかった。PN と比べ、OC に対する自動係り受け解析結果は頑健ではない可能性があるが、これが述語項構造解析にどのような影響を与えているかは、引き続き調査する必要がある。



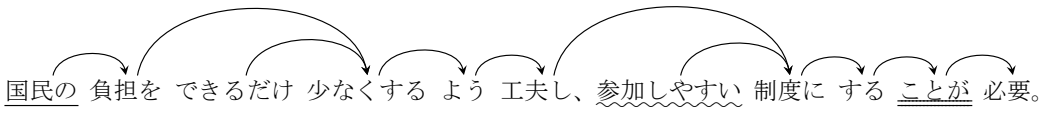
■述語と項が係り受け関係にない場合　以下の事例は訓練と評価にともに新聞記事を用いた事例である。述語と項が離れているが、PN 分野では人名である場合項になりやすいという傾向を学習しているため、文頭の人物を捉えられている。



次の事例も訓練と評価にともに新聞記事を用いた事例であるが、正解を出力できなかった。係り受け関係にある表層格がガ格である近くの句を選んでしまったと考えられる。自動係り受け解析に失敗しているために項同定も失敗している可能性があるので、今後は人手でアノテーションされた係り受け情報[†6]を用いた実験を試したい。



次の事例もまた訓練と評価にともに新聞記事を用いた事例であるが、正解を出力できなかった。「参加する」という述語と「国民」「こと」それぞれの選択選好に関する素性を用いることで、正解を選べるようになる可能性があると考えられる。



6　おわりに

本稿では、複数の分野のテキストで機械学習に基づく述語項構造解析器を訓練し、分野の異なるテキストでの性能を評価した。評価の結果、分野の異なるテキストを訓練に用いた場合、性格の異なるモデルを学習することが分かった。また、訓練と評価に用いるデータの分野が同じ場合が最も文内ガ格の項同定精度が高いことが確認できた。一方、訓練に新聞記事と Yahoo! 知恵袋のデータを両方用いる効果は限定的で、特に評価に新聞記事を用いた場合、Yahoo!知恵袋のデータを新聞記事に加えて訓練すると、係り受け関係にある場合もない場合も大きく精度が下がることが分かった。

今後は共起素性など意味的な素性、そして表層に関する素性を入れた場合にどのように結果

[†6] https://sites.google.com/site/masayua/bccwjdep

が変わるか調査するとともに、白書や書籍といった他の分野のテキストにおいても分野の影響がどのようにあるのか検討してみたい。

参考文献

[Ima09] Imamura, K., K. Saito, and T. Izumi: Discriminative Approach to Predicate-Argument Structure Analaysis with Zero-Anaphora Resolution, in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 85–88, 2009.

[Kaw02] Kawahara, D., S. Kurohashi, and K. Hasida: Construction of a Japanese Relevance-tagged Corpus, in *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pp. 2008–2013, 2002.

[Tai08] Taira, H., S. Fujita, and M. Nagata: A Japanese Predicate Argument Structure Analysis using Decision Lists, in *Proceedings of EMNLP-2008*, pp. 523–532, 2008.

[Yos11] Yoshikawa, K., M. Asahara, and Y. Matsumoto: Jointly Extracting Japanese Predicate-Argument Relation with Markov Logic, in *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pp. 1125–1133, 2011.

[橋本 11] 橋本, 黒橋, 河原, 新里, 永田：構文・照応・評判情報つきブログコーパスの構築, 自然言語処理, Vol. 18, No. 2, pp. 175–201, 2011.

[笹野 11] 笹野, 黒橋：大規模格フレームを用いた識別モデルに基づく日本語ゼロ照応解析, 情報処理学会論文誌, Vol. 52, No. 12, pp. 3328–3337, 2011.

[小町 11] 小町, 飯田：BCCWJ に対する述語項構造と照応関係のアノテーション, 日本語コーパス平成 22 年度公開ワークショップ, pp. 325–330, 2011.

[小椋 09] 小椋, 小木曽, 小磯, 冨士池, 宮内, 渡部, 竹内, 小川, 小西, 原, 中村：「『現代日本語書き言葉均衡コーパス』における形態論情報付与作業の進捗状況」,『特定領域「日本語コーパス」平成 20 年度公開ワークショップ予稿集』, pp. 57–64, 2009.

[飯田 04] 飯田, 乾, 松本：文脈的手がかりを考慮した機械学習による日本語ゼロ代名詞の先行詞同定, 情報処理学会論文誌, Vol. 45, No. 3, pp. 906–918, 2004.

[飯田 10] 飯田, 小町, 井之上, 乾, 松本：述語項構造と照応関係のアノテーション: NAIST テキストコーパス構築の経験から, 自然言語処理, Vol. 17, No. 2, pp. 25–50, 2010.