# ビジネスメール文に対する日本語述語項構造解析の検討

平 博順　　田中 貴秋　　藤田 早苗　　永田 昌明
NTT コミュニケーション科学基礎研究所
{taira.hirotoshi, tanaka.takaaki, fujita.sanae, nagata.masaaki}@lab.ntt.co.jp

## 1　はじめに

入力文中の述語および項，さらにそれらの役割を特定する述語項構造解析技術は，機械翻訳やテキストマイニングを行う上での基盤技術となっている [3, 4, 7, 8, 9, 10, 16, 17, 18]．特に主語，目的語等の省略が他の言語に比べて頻繁に起こる日本語述語項構造解析は，日本語と他言語との間の統計的機械翻訳 [1, 12] や，日本語テキストからのマイニングなどでその重要性が高まっている．ところで，日本語の述語項構造解析の研究で扱われているコーパスは新聞記事に対するものが中心であった．しかし，最近では，書籍，白書，質問サイトも対象としたもの（現代日本語書き言葉均衡コーパス (BCCWJ)） [5, 14] やブログを対象としたもの（京大 NTT ブログコーパス (KNB)） [2] も構築され，さまざまなドメインに対する述語項構造解析について研究が進められている．

本稿では，これまでほとんど扱われてきていない，ビジネスメール文についての述語項構造解析について検討を行う．社内外のビジネスパートナーと様々なやりとりを行うビジネスメール文については，他の対象と比べ，省略表現が多い，定型文が多い，敬語表現が多い，などの異なる特徴があることが想定される．本稿では，ビジネスメールの文例集のデータを基に，人手で述語項構造のアノテーションを行い，このデータに対して，新聞記事で訓練した日本語述語項構造解析器で解析を行い，新聞記事を解析する場合との差異について検討を行った．

本稿の構成は，以下の通りである．まず，2 章において，本研究で用いたビジネスメール文例データについて，簡単に説明する．3 章で，本研究で用いた述語項構造解析器について説明し，4 章において，評価実験ついて述べ，最後にまとめを行う．

表 1: 実験に使用したビジネスメール文例データの一部

| |
|---|
| (CL ビジ) |
| ・下記まで貴社についての資料をお送りいただければ幸いです. |
| ・昨日，当社のカタログを郵送いたしました. |
| ・締切まであまり時間がなくてすみません. |
| ・喜んでお手伝いいたします. |
| ・あなたの考えをお聞かせください. |
| (SW ビジ) |
| ・一週間以内に打ち合わせをして署名してもらえませんか？ |
| ・早急に山田部長にご連絡したいことがございます. |
| ・正直な感想を聞かせてください. |
| ・翌日配送を選択したらいつ届きますか？ |
| ・またすぐに必ず私からご連絡します. |

## 2　ビジネスメール文例データ

ビジネスメールを公開できる形で入手するのは，企業秘密の観点から難しいため，今回は実際のビジネスメールの代替として，市販されている下記のビジネスメールの文例集のデータを用いた．

- 日英ビジネス文対訳データ（クロスランゲージ社）（以下「CL ビジ」と略す）
- 大文嶺：テキスト対訳・ビジネスメール分野・米語=日本語・2012 年版（ストレートワード社）（以下「SW ビジ」と略す）

前者のデータは，元々翻訳メモリ用の日英対訳データであり，後者のデータは，日本語話者が英語でビジネスメールを書く場合の文例集である．表 1 に各データの例を示す．どちらのデータも，商品のやりとりについての内容のメールや，社内でのメールのやりとり，といったものを含んだ内容になっている．また，文同士に関連は無く，それぞれ独立した 1 文単位の

表 2: 項正解の内訳

| | ガ格項 | | | | | |
|---|---|---|---|---|---|---|
| 位置タイプ | NTC14b | | CL ビジ | | SW ビジ | |
| DEP | 57,049 | (53.9%) | 775 | (56.4%) | 412 | (44.7%) |
| SAME_BS | 209 | (0.1%) | 8 | ( 0.5%) | 5 | ( 0.5%) |
| INTRA_Z | 19,580 | (18.4%) | 130 | ( 9.4%) | 20 | ( 2.1%) |
| INTER_Z | 13,093 | (12.3%) | - | - | - | - |
| EXO1 | 2,517 | (2.3%) | 231 | (16.8%) | 286 | (31.0%) |
| EXO2 | 133 | (0.1%) | 75 | ( 5.4%) | 140 | ( 8.1%) |
| EXOG | 15,907 | (15.0%) | 155 | (11.2%) | 57 | ( 6.1%) |
| 全体 | 105,838 | (100.0%) | 1374 | (100.0%) | 920 | (100.0%) |
| | ヲ格項 | | | | | |
| 位置タイプ | NTC14b | | CL ビジ | | SW ビジ | |
| DEP | 38,190 | (88.8%) | 401 | (68.9%) | 283 | (78.8%) |
| SAME_BS | 95 | (0.2%) | 19 | ( 3.2%) | 12 | ( 3.3%) |
| INTRA_Z | 3,301 | (7.6%) | 130 | (22.3%) | 20 | ( 5.5%) |
| INTER_Z | 1,299 | (3.0%) | - | - | - | - |
| EXO1 | 13 | (0.03%) | 2 | ( 0.3%) | 1 | ( 0.2%) |
| EXO2 | 9 | (0.02%) | 0 | ( 0.0%) | 2 | ( 0.5%) |
| EXOG | 74 | (0.1%) | 30 | ( 5.1%) | 41 | (11.4%) |
| 全体 | 42,981 | (100.0%) | 582 | (100.0%) | 359 | (100.0%) |
| | ニ格項 | | | | | |
| 位置タイプ | NTC14b | | CL ビジ | | SW ビジ | |
| DEP | 19,152 | (89.0%) | 136 | (70.8%) | 54 | (53.4%) |
| SAME_BS | 702 | (3.2%) | 14 | ( 7.2%) | 8 | ( 7.9%) |
| INTRA_Z | 1,076 | (5.0%) | 1 | ( 0.5%) | 2 | ( 1.9%) |
| INTER_Z | 540 | (2.5%) | - | - | - | - |
| EXO1 | 10 | (0.04%) | 5 | ( 2.6%) | 14 | (13.8%) |
| EXO2 | 3 | (0.01%) | 15 | ( 7.8%) | 12 | (11.8%) |
| EXOG | 32 | (0.1%) | 21 | (10.9%) | 11 | (10.8%) |
| 全体 | 21,515 | (100.0%) | 192 | (100.0%) | 101 | (100.0%) |

データである．なお，実験には対訳データの日本語側のテキストのみを使用した．これら 2 種類のデータからそれぞれ，1018 文，859 文について，NAIST テキストコーパスの仕様に準じて人手で述語項構造のアノテーションを行った．その結果，項の種類毎の内訳は，表 2 のようになった．ここで，位置タイプとは，述語と項（ガ格，ヲ格，ニ格）との間の，係り受け状態，および同一文，同一文節にあるか否か，外界照応であるかを示すものである．本稿では，述語と項の間に係り受け関係があるものを DEP，同一文節にあるものを SAME_BS，DEP でも SAME_BS でもないが，同一文中にある場合を INTRA_Z，述語と項とが異なる文にある場合を INTER_Z，項が外界照応の関係にあって 1 人称を指すものを EXO1，2 人称を指すものを EXO2，それ以外の外界照応を EXOG として表している．

この表から，新聞記事のデータである NAIST テキストコーパス Ver1.4b (NTC14b) と比べて，「CL ビジ」「SW ビジ」共に，外界照応の割合がかなり高いことが見て取れる．

## 3 述語項構造解析器

本稿では，述語項構造解析器として，新聞記事データを用いた実験で解析精度の高かった拡張対立候補モデル [11] の方法を用いている．この方法は，Twin Candidate モデル [13] と，位置タイプ別分類を組み合わせた方法である．図 1 示した処理手順のように，まず，述語項の候補を DEP，INTRA_Z といった位置タイプで区別し，位置タイプごとの分類器で最も適切と思われる述語項の候補を選ぶ．さらに各位置タイプ別解析器の 1 位同士で総合 1 位を選ぶ拡張対立候補モデルの解析器で，最終的な述語項を選ぶ．ただし，どの述語項の候補も，適切ではないと判定された場合には，外界照応とみなされ，さらに人称分類器で，EXO1，EXO2，EXOG のいずれかに分類される．なお，本稿の実験では，学習器として LIBINEAR (Ver. 1.92) のロジスティック回帰を使用し，パラメータはデフォルト値で実験を行った．

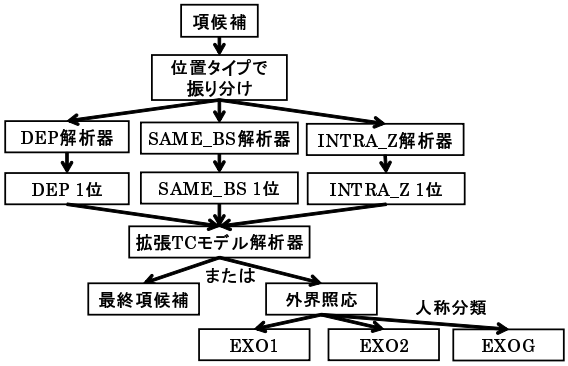


図 1: 解析の流れ

# 4 評価実験

実験では，NAIST テキストコーパス 1.4$\beta$ を学習データとして述語項構造解析器の解析モデルを学習した後,「CL ビジ」「SW ビジ」データを評価データとして解析し，その精度を評価した．

また，項の種類については，NAIST テキストコーパスの場合に準じて，ガ格，ヲ格，二格の三種類の項について評価を行った．

## 4.1 学習に使用した特徴量

解析モデルの学習には，[11] で用いられている特徴量を用い，以下に示す大きく分けて 4 つのタイプの特徴量を使用した．

- 対象述語（PRED）に対する特徴量
  対象述語の語彙，品詞，態，述語の語尾の機能表現，疑問代名詞の有無，など．
- 項候補（NP）に対する特徴量項候補の語彙，品詞，固有表現，代名詞の分類，後続する助詞，項候補の文書中の出現位置など．
- 項候補と対象述語との間の関係に関する特徴量項候補と対象述語の間の係り受け関係，隣接関係，語彙の組み合わせ，河原らの Web コーパス [15] での格関係出現有無，項候補と対象述語間の距離，など．
- 文脈に関する特徴量文章中で焦点となっている語への成り易さの一つの指標となる Salient Reference List [6] に基づくスコア．

## 4.2 実験結果

精度の評価は，述語とそれに対する項スロットはあらかじめ与えられているとし，各項のスロットに対する項のシステム出力について，精度 (Accuracy) による評価を行った．図 3 にその結果を示す．

表 3: 解析精度 (単位:%)

| | ガ格項 | |
|---|---|---|
| 位置タイプ | CL ビジ | SW ビジ |
| DEP | 77.2 | 68.1 |
| SAME_BS | 0.0 | 25.0 |
| INTRA_Z | 51.5 | 62.6 |
| EXO1 | 43.8 | 0.0 |
| EXO2 | 0.0 | 0.0 |
| EXOG | 14.6 | 7.1 |
| 全体 | 56.6 | 44.9 |

| | ヲ格項 | |
|---|---|---|
| 位置タイプ | CL ビジ | SW ビジ |
| DEP | 97.7 | 91.3 |
| SAME_BS | 66.6 | 0.0 |
| INTRA_Z | 2.0 | 5.1 |
| EXO1 | 0.0 | 0.0 |
| EXO2 | 0.0 | 0.0 |
| EXOG | 35.0 | 7.1 |
| 全体 | 81.7 | 76.6 |

| | 二格項 | |
|---|---|---|
| 位置タイプ | CL ビジ | SW ビジ |
| DEP | 86.4 | 69.8 |
| SAME_BS | 100.0 | 80.0 |
| INTRA_Z | 0.0 | 0.0 |
| EXO1 | 0.0 | 0.0 |
| EXO2 | 0.0 | 0.0 |
| EXOG | 30.3 | 35.7 |
| 全体 | 77.0 | 63.3 |

これらの結果を見ると，特に外界照応のヲ格項，二格項の EXO1，EXO2 については，全く解析できていないことが分かる．これは，学習データに使用した NTC14b コーパスに，これらのデータがほとんど含まれておらず，適切な学習ができていないのが原因だと考えられる．表 4 に解析が誤った具体例を示す．pred が述語位置を表し，gold が述語に対する項の人手正解位置，sys が項のシステム出力位置を表す．これを見ると，敬語に関する表現が，格フレーム辞書に登録されていないことによる解析誤りが多く発生していることが分かる．また，複合語について，述語と項の定義が曖昧であることに起因する問題もあることも分かった．

表 4: 解析誤りの例

あなた $_{sys,ga}$/に/は /関係 $_{gold,ga}$/あり $_{pred}$
/ませ/ん/から/.
(※ SAME_BS, 複合語の扱い)

遅延/を お許し/いただけ/ます よう お願い $_{gold,o}$
/申し上げ $_{pred}$/ます/. (sys: exog)
(※ SAME_BS, 「申し上げる」の述語項の定義)

お/役/に 立てる こと/が ござい/まし/たら
お知らせ $_{pred}$/ください/. (gold,ni: EXO1, sys: $\phi$)
(※ EXO1, 「お知らせ」が格フレーム辞書に無かった)

貴社/について/の 資料/を お送り $_{pred}$
/いただけれ/ば 幸い/です/.
(gold,ga: EXO2, sys,ga: $\phi$)
(※ EXO2, 「お送りいただく」の言い回しが
考慮されていない)

到着/ロビー/で お/出迎え $_{pred}$/いたし/ます/.
(gold,o: exo2, sys: $\phi$)
(※ EXO2, 「お出迎え」の言い回しが述語として
考慮されていない)

## 5 おわりに

本稿では，新聞記事データで訓練した述語項構造解析器を用いて，ビジネスメール文についての日本語述語項構造解析を行い，どの程度，解析に使用可能かについて簡単な実験で調べた．その結果，係り受け関係にある述語項や，同一文内にある述語項の場合は，比較的流用できる可能性があるものの，やはり外界照応の関係については，データを増やしてモデルを作り直す必要があることが示唆された．また，敬語表現についてこれまで特徴量として考慮していないが，敬語表現について学習段階で考慮した方がよいことが伺えた．

## 参考文献


[1] 古市将仁, 村上仁一, 徳久雅人, 村田真樹. 日英統計翻訳における主語補完の効果. 言語処理学会 第 17 回年次大会 発表論文集, pp. 163–166, 2011.

[2] 橋本力, 黒橋禎夫, 河原大輔, 新里圭司, 永田昌明. 構文・照応・評価情報つきブログコーパスの構築. Vol. 18, No. 2, pp. 175–201, 2011.

[3] Yuta Hayashibe, Mamoru Komachi, and Yuji Matsumoto. Japanese predicate argument structure analysis exploiting argument position and type. pp. 201–209, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing.

[4] Kenji Imamura, Kuniko Saito, and Tomoko Izumi. Discriminative approach to predicate-argument structure analysis with zero-anaphora resolution. In *Proc. of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 85–88, 2009.

[5] 小町守, 飯田龍. 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.

[6] S. Nariyama. *Ellipsis and reference tracking in Japanese*, Vol. 66. John Benjamins Publishing Company, 2003.

[7] R. Sasano, D. Kawahara, and S. Kurohashi. A fully-lexicalized probabilistic model for japanese zero anaphora resolution. In *Proc. of COLING*, Vol. 8, pp. 769–776, 2008.

[8] Hirotoshi Taira, Sanae Fujita, and Masaaki Nagata. A japanese predicate argument structure analysis using decision lists. In *Proc. of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2008.

[9] Hirotoshi Taira, Sanae Fujita, and Masaaki Nagata. Predicate argument structure analysis using transformation based learning. In *Proc. of the Conference on ACL 2010*, 2010.

[10] 平博順, 永田昌明. 構造学習を用いた述語項構造解析. 言語処理学会 第 14 回年次大会, pp. 556–559, 2008.

[11] 平博順, 永田昌明. 述語項構造解析を伴った日本語省略解析の検討. 言語処理学会 第 19 回年次大会, pp. 106–109, 2013.

[12] 平博順, 須藤克仁, 永田昌明. 統計翻訳における日本語省略補完の効果の分析. 言語処理学会 第 18 回年次大会, pp. 135–138, 2012.

[13] X. Yang, J. Su, and C.L. Tan. A twin-candidate model for learning-based anaphora resolution. *Computational Linguistics*, Vol. 34, No. 3, pp. 327–356, 2008.

[14] 吉本暁文, 小町守, 松本裕治. 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－. 第 3 回コーパス日本語学ワークショップ, 2013.

[15] 河原大輔, 黒橋禎夫. 高性能計算環境を用いた web からの大規模格フレーム構築. 情報処理学会 自然言語処理研究会, pp. 67–73, 2006.

[16] 河原大輔, 黒橋禎夫. 自動構築した大規模格フレームに基づく構文・格解析の統合的確率モデル. 自然言語処理, Vol. 14, No. 4, pp. 67–81, 2007.

[17] 吉川克正, 浅原正幸, 松本裕治. Markov logic による日本語述語項構造解析. 情報処理学会研究報告（自然言語処理研究会）2010-NL-199 No.5, 2010.

[18] 渡邉陽太郎, 浅原正幸, 松本裕治. 述語語義と意味役割の結合学習のための構造予測モデル. 人工知能学会論文誌, Vol. 25, No. 2, pp. 252–261, 2010.