述語項構造解析を伴った日本語省略解析の検討

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

1　はじめに

日本語では，主語，目的語等の省略が頻繁に起こる．この日本語の省略を高精度に自動補完する技術は，日本語と他言語との間の機械翻訳 [3, 15] や，日本語テキストからのマイニングなどで重要な技術である．

これまで，日本語の省略解析については多くの従来研究がある [1, 2, 9, 18, 10]．しかし，大規模コーパスを対象として評価を行った研究はそれほど多くはない．

大規模コーパスを対象とした研究では，山本らが，日本語の話し言葉を対象に学習ベースでの研究がある [16]．この研究では，省略されている箇所が与えられたときに，その部分が外界照応であるか，それとも文内に先行詞があるかの分類を行い，外界照応と分類された場合は，さらにその人称，単複の分類を行っている．

本稿では，これを発展させ，与えられた文内に省略があるかどうかが与えられていない条件の下で，省略があるかどうかの判定，省略があった場合，文内ゼロ代名詞，文間ゼロ代名詞の先行詞が同一テキスト内にあれば，それを具体的に特定し，外界照応と判定されれば，その人称の分類を行う解析器の作成を試みた．

このような解析器は，日本語の広義の述語項構造解析器を拡張したものと見ることができる．このような視点から，大規模な日本語述語項構造コーパスである NAIST テキストコーパスを対象にして，この述語項構造 [4, 8] の解析および外界照応の判定と分類を組み合わせることで，これが実現できるかどうか，検討を行った．

日本語の述語構造解析については，これまで多くの研究がなされてきている [5, 7, 11, 12, 13, 14, 20, 21, 22]．本稿ではその中で，林部らが行っているトーナメントモデルを用いる方法 [5] をベースに，外界照応の解析も合わせて行う解析器を作成し，大規模コーパスである NAIST テキストコーパス [23] を対象に評価を行った．

本稿の構成は，以下の通りである．まず，2 章において，本研究で試みた，従来の述語項構造解析を省略解析まで拡張した方法について述べる．3 章において，評価実験の設定について述べ，5 章で，実験結果について述べ，最後にまとめを行う．

2　述語項構造解析を伴った省略解析方法の検討

2.1　本稿で扱う解析の範囲

NAIST テキストコーパスを対象に述語項構造解析を行う際，従来研究では，問題の性質の違いから，次の 3 種類の項を区別して解析を行うことが多かった．

- 述語を含む文節と，項を含む文節が，係り受け関係にある場合（DEP）
- 述語を含む文節と，項を含む文節が，係り受け関係にはないが，同一文中にある場合 (文内ゼロ代名詞の先行詞)（INTRA_Z）
- 述語を含む文節と，項を含む文節が，異なる文にある場合 (文間ゼロ代名詞の先行詞)(INTER_Z)

これら 3 種類の項に加えて，我々は，

- 述語を含む文節と，項を含む文節が同じ文節にある場合 (SAME_BS)

および，

- 項が外界照応 (Exophora) である場合 (EXO)

を考慮して解析を行う．

ここで，一般に項が「省略されている」状態というのは，INTRA_Z, INTER_Z, EXO の項が正解である場合を指すと考えることができる．

また，外界照応に関しては，それが指すものについての人称の情報は，機械翻訳，マイニングなどの応用

では重要である．そこで，それらの特定も包含した形で解析を行うこととする．

実際に応用を考えた場合，省略解析では，次の 3 段階の処理を行う必要がある．

1. 述語の格フレームを特定し，必須の項が何であるかを判定する
2. 1) で分かった埋めるべき項について，述語項構造解析を行い，その項が，同一テキスト中に出現しているかどうかを解析する．
3. 2) で項が同一テキスト中に出現していない場合，項が省略されていると判定し，さらに省略項のタイプ（人称，性別，単複等）を特定する

ここで，1) に関しては，述語に対する格フレームの辞書を規定した上で，その格フレームのうちのどれにあてはまるかの曖昧性解消を行うことが本来必要であるが，今回扱った NAIST テキストコーパスでは明示的にそのような辞書が規定されていないため，今回は，実際の答えの項のスロットは既に決まっているものとして扱った．

また，2) に関しては，文内，文間ゼロ代名詞の先行詞特定も含めた形での述語項構造解析を行った．

3) では，省略項のタイプを特定するが，NAIST テキストコーパスでは，外界照応のタグは EXO1，EXO2，EXOG の 3 種類の付与のみなので，この分類を行うところまでを解析対象とした．

また，項のタイプは NAIST テキストコーパスではガ格, ヲ格, 二格の 3 種類が項の対象となっており，これら 3 種類を解析することとする．

2.2 解析方法の概要

従来の述語項構造解析器の部分については，林部らのトーナメントモデル [6, 17] を用いる方法などベースに拡張し，図 1 のような処理手順を試みた．

まず，

(a) 形態素解析器および係り受け解析器（本稿では Cabocha (Ver. 0.60pre) を使用）を用いて，テキストを解析．得られた形態素のうち，項候補になりにくい機能語等を項候補分類器を使ってフィルタリングする．

(b) 項候補で，項候補と解析対象の述語が文節レベルで係り受け関係になる場合を DEP，同一文節にある場合を SAME_BS，係り受け関係になく，同一文節でもないが，同一文内にある場合を INTRA_Z，同一テキスト内だが，異なる文にある場合を INTER_Z として，それぞれで，最も項になりやすい項候補を分類器を使って選出し、代表の項候補とする．

(c) SAME_BS の代表と INTRA_Z の代表でどちらが項となりやすいかを分類

(d) DEP の代表と (c) の勝者でどちらが項となりやすいかを分類

(e) INTER_Z と (d) の勝者でどちらが項となりやすいかを分類

(f) (e) の勝者 (NO_EXO) か，外界照応 (EXO) のどちらが当てはまりやすいかを分類

(g) (f) で外界照応と分類された場合は，人称判定 (1 人称 (EXO1), 2 人称 (EXO2), 総称 (EXOG)) を分類する．

なお，学習器には 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．

3 評価実験の設定

3.1 実験対象コーパス

実験対象のコーパスには，NAIST テキストコーパス 1.4β を用いた．NAIST テキストコーパスでは，一般的な述語と事態性名詞についてのアノテーションがなされているが，今回は一般的な述語を対象に実験を行った．

NAIST テキストコーパス中の述語（pred）と項（ガ格, ヲ格, 二格）の組み合わせについて，項の分類ごとの分布を表 1 に示す．

実験では，これら全部の項について実験対象とした．

3.2 学習に使用した特徴量

解析モデルの学習には，従来の研究 [5, 7, 12, 16] で用いられている特徴量を参考に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．

3.2.1 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．

3.2.2 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．

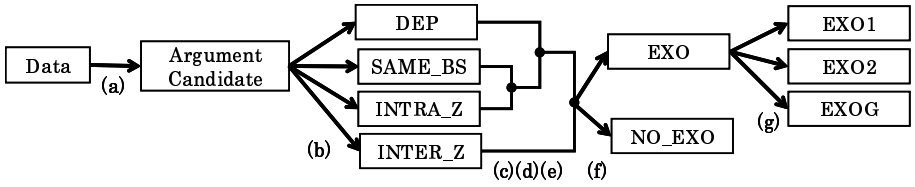

図 1: 解析の流れ

表 1: 用いられた述語の項タイプの分布

	述語の項タイプ							
	ガ格項		ヲ格項		二格項		合計	
文内係り受け有 (DEP)	57,049	(53.9%)	38,190	(88.8%)	19,152	(89.0%)	89,093	(61.9%)
同一文節 (SAME_BS)	209	(0.1%)	95	(0.2%)	702	(3.2%)	1,005	(0.6%)
文内係り受け無 (INTRA_Z)	19,580	(18.4%)	3,301	(7.6%)	1,076	(5.0%)	23,060	(16.0%)
文間ゼロ代名詞 (INTER_Z)	13,093	(12.3%)	1,299	(3.0%)	540	(2.5%)	14,559	(10.1%)
外界照応以外の項全体	89,931	(84.9%)	42,885	(99.7%)	21,470	(99.7%)	127,717	(88.8%)
1 人称 (EXO1)	2,517	(2.3%)	13	(0.03%)	10	(0.04%)	2,540	(1.7%)
2 人称 (EXO2)	133	(0.1%)	9	(0.02%)	3	(0.01%)	145	(0.1%)
総称 (EXOG)	13,257	(12.5%)	74	(0.1%)	32	(0.1%)	13,329	(9.2%)
外界照応 (EXO) 全体	15,907	(15.0%)	96	(0.2%)	45	(0.2%)	16,014	(11.1%)
項スロット全体	105,838	(100.0%)	42,981	(100.0%)	21,515	(100.0%)	143,731	(100.0%)

3.2.3 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 [19] での格関係出現有無，項候補と対象述語間の距離，など．

3.2.4 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 [10] に基づくスコア．

4 実験結果

精度の評価は，記事単位での 5-fold cross validation で行った．なお精度は，非外界照応の項，外界照応の項を分け，それぞれの適合率 (Precision)，再現率 (Recall)，F_1 値を用いた．

図 2 にその結果を示す．なお，EXO1 のヲ，二格，EXO2 については，学習時に訓練サンプルがほとんど得られず学習できなかったため省略している．係り受け関係のある DEP については，ガ，ヲ，二格どれも 7 割を超える F 値を示したが，それ以外の係り受け関係にないものについては，同一文節の二格以外は，精度は悪かった．分析したところ，EXO と NO-EXO の判定を行う際に，現状の分類器では，より外界照応と判定する傾向にあり，そのために，INTRA_Z，INTER_Z の精度が，低下していることが分かった．

一方，外界照応に関しては，総称の分類精度が割合高かったものの，1 人称については精度が低かった．これは外界照応の判定で使用する特徴量，データセットの量が十分でなかったためである可能性が高い．

5 おわりに

本稿では，日本語の述語項構造解析と省略解析を同時に解析する解析方法について検討し，評価実験を行った．精度については，述語項構造解析についても省略解析についても改善の余地があり，今後，処理方法の改良や，使用特徴量，訓練データの追加により精度向上を図っていきたいと考えている．

参考文献

[1] C. Aone and S.W. Bennett. Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pp. 122–129. Association for Computational Linguistics, 1995.

表 2: 実験結果 (単位:%)

項タイプ	ガ格項			ヲ格項			二格項			合計		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
DEP	80.5	67.1	73.2	90.9	69.5	78.7	93.5	57.3	71.1	85.6	66.3	74.7
SAME_BS	66.7	0.98	1.93	100.0	17.9	30.4	97.1	53.0	68.6	97.0	39.1	55.7
INTRA_Z	63.7	0.26	0.51	76.9	0.30	0.60	25.0	0.09	0.18	63.9	0.25	0.51
INTER_Z	15.2	0.46	0.90	0.00	0.00	0.00	25.0	0.18	0.36	15.1	0.41	0.80
外界照応以外の項全体	79.9	42.7	55.7	90.8	61.9	73.6	93.6	52.9	67.6	85.3	49.5	62.6
EXO1	40.9	24.3	30.5	-	-	-	-	-	-	40.9	24.1	30.3
EXO2	-	-	-	-	-	-	-	-	-	-	-	-
EXOG	19.2	81.5	31.1	0.37	70.3	0.75	0.27	81.2	0.55	13.7	81.4	23.4
外界照応 (EXO) 全体	19.7	71.7	31.0	0.37	54.2	0.75	0.27	57.8	0.55	14.2	71.6	23.7

[2] Kohji Dohsaka. Identifying the referents of zero-pronouns in japanese based on pragmatic constraint interpretation. In *Proceedings of ECAI*, pp. 240–245, 1990.

[3] 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.

[4] J. Grimshaw. *Argument structure*. the MIT Press, 1990.

[5] 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.

[6] R. Iida, K. Inui, H. Takamura, and Y. Matsumoto. Incorporating contextual cues in trainable models for coreference resolution. In *Proceedings of the 10th EACL Workshop on the Computational Treatment of Anaphora*, pp. 23–30. Citeseer, 2003.

[7] 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.

[8] 影山太郎. 動詞意味論－言語と認知の接点－. くろしお出版, 1997.

[9] H. Nakaiwa and S. Yamada. Automatic identification of zero pronouns and their antecedents within aligned sentence pairs. In *Proc. of the 3rd Annual Meeting of the Association for Natural Language Processing*, 1997.

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

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 山本和英, 隅田英一郎. 決定木学習による日本語対話文の格要素省略補完. 自然言語処理, Vol. 6, No. 1, pp. 3–28, 1999.

[17] 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.

[18] K. Yoshimoto. Identifying zero pronouns in japanese dialogue. In *Proceedings of the 12th conference on Computational linguistics-Volume 2*, pp. 779–784. Association for Computational Linguistics, 1988.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 飯田龍, 小町守, 乾健太郎, 松本裕治. NAIST テキストコーパス: 述語項構造と共参照関係のアノテーション. 情報処理学会研究報告 (自然言語処理研究会) NL-177-10, pp. 71–78, 2007.