分散表現による大規模格フレームの汎化を利用した統合的ニューラルゼロ照応解析

山城 颯太　西川 仁　徳永 健伸
東京工業大学 情報理工学院
{yamashiro.s.aa@m, {hitoshi, take}@c}.titech.ac.jp

1 はじめに

日本語ゼロ照応解析においては，大規模な訓練事例を利用した文間ゼロ照応および文内ゼロ照応を統一的に解決するモデルが存在しない．本論文はこれらの要件を満たす，より頑健な統一的ゼロ照応解析モデルを提案する．

文内，文間のゼロ照応解析を同時に行っている既存研究として Sasano and Kurohashi (2011) が挙げられるが，これは学習に Web コーパスのみを使用しておりデータ量も少ない．従来のゼロ照応解析研究の多くは，NAIST Text Corpus (NTC) (Iida et al., 2007) と呼ばれる新聞コーパスを使用し，文内ゼロ照応のみに焦点を絞っている (Shibata et al., 2016; Ouchi et al., 2017; Matsubayashi and Inui, 2017)．しかしゼロ照応解析結果の応用を考えた時，むしろアプリケーションの全体性能を大幅に悪化させるのは文間ゼロ照応であり，また解析対象の文書ドメインについても，新聞のみならず Blog，QA，書籍，白書，雑誌などあらゆるドメインの文書に対して頑健なゼロ照応解析手法こそより有用性が高い．

一方で，文内，文間のゼロ照応解析を同時に行う際，正例と負例の比率が約 1 対 1000 と著しく不均衡となる上，解析の対象となる名詞が大幅に増加する．これらは計算量を大幅に増幅させ，かつモデルの汎化を妨げる要因となる．

本研究では大規模格フレームを利用したゼロ照応解析において，分散表現と格フレームに含まれる項の平均ベクトルを用いた効率的な候補削減手法を取り入れることで様々なドメインの文書への対応を可能とし，より汎用性の高い文内，文間の同時ゼロ照応解析手法を提案する．なお，BCCWJ 全体を用いた文内文間のガヲニ格を対象とするゼロ照応解析は．本研究が初の試みである．

2 関連研究

2.1 ゼロ照応解析

Matsubayashi and Inui (2017) はフィードフォワードニューラルネットワーク (FNN) を用いて，NTC に対して直接の係り受け関係と文内のゼロ照応解析を同時に行っている． Sasano and Kurohashi (2011) は対数線形モデルを用いて，Web コーパスに対して文内と文間のゼロ照応解析を同時に行っている．これらに対して我々はランキング SVM[1] (Joachims, 2006) を用いて,『現代日本語書き言葉均衡コーパス』 (BCCWJ) (Maekawa et al., 2014) に対して，文内，文間のゼロ照応解析を同時に行う．

2.2 大規模格フレーム

ゼロ照応解析のための有用な言語資源として格フレームがある．格フレームとは述語とその述語が取りうる項を述語の格パターンごと，格ごとに整理した共起情報である．例えば「オープンしてる」のガ格には「店」などの施設を示す名詞が入りやすく，ニ格には「近く」などの場所を示す名詞が入りやすいと考えられる．

しかし一方で，ガ格に「サイト」や「ページ」などの名詞が入った場合には，ニ格には格要素の入らない可能性が高い (照応なし)．これらを「オープンしてる」についての別の格パターンとして分けておくことで，述語や格要素間の語彙的選好の知識を照応解析に利用することができる (Sasano et al., 2008; Sasano and Kurohashi, 2011; Hangyo et al., 2013)．格フレームの構築に関しては Kawahara and Kurohashi (2006) が Web テキストから格フレームを自動構築する手法を提案している．これらの大規模 Web コーパスから取得，整理された格フレーム知識は図 1 のように京大格フレーム[2]として公開されている．

3 解析モデル

3.1 ベースモデル

本研究の提案手法は素性の設計に関して，Hangyo et al. (2013) のモデルをベースとしている．これについて説明する．

[1] https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

[2] http://www.gsk.or.jp/catalog/gsk2008-b/ ただし，リンク先の京大格フレームは古い版であり，本項において使用したものは未公開の新しい版である．

表 1: 「オープンする:動 1」の格フレーム

格	格要素	出現回数	ベクトル
ガ	店	129	$\phi_{店}$
	カフェ	38	$\phi_{カフェ}$
	ショップ	37	$\phi_{ショップ}$
	...	...	...
	合計	231	
ニ	近く	6	$\phi_{近く}$
	跡地	2	$\phi_{跡地}$
	ところ	2	$\phi_{ところ}$
	...	...	...
	合計	57	

3.1.1 素性による述語項構造の表現

入力テキスト (t) の解析対象述語 (p) に格フレーム (cf) を割り当てる．その格フレームの格スロットのうち，直接係り受け関係にある名詞句と対応付けられなかった格スロットに対応する格要素がゼロ代名詞になると仮定し，各格スロットと名詞句の対応付けを (a) とした述語項構造を表現する素性ベクトル $\phi(cf, a, p, t)$ を以下のように 4 つのベクトル結合として表現する．

$$\begin{aligned}\phi(cf, a, p, t) \quad = \quad & (\phi_{overt}(cf, a_{overt}), \\ & \phi_{case}(cf, ガ \leftarrow e_{ガ}), \\ & \phi_{case}(cf, ヲ \leftarrow e_{ヲ}), \\ & \phi_{case}(cf, ニ \leftarrow e_{ニ}))\end{aligned}$$

ここで $\phi_{overt}(cf, a_{overt})$ は直接係り受けがある述語項構造を表わす素性ベクトルであり，コーパスから算出された表層格生成確率と入力表層格の種類，京大格フレームから算出された格スロット生成確率から計算される．$\phi_{case}(cf, c \leftarrow e)$ は格 c に先行詞 e が割り当てられることを表わす素性ベクトルである．格 c に先行詞 e が対応付けられない「照応なし」,「外界ゼロ照応」の場合，各格に対応する素性ベクトル $\phi_{case}(cf, c \leftarrow e)$ は cf と c のみに依存する素性以外をすべてゼロとする．$\phi_{overt}(cf, a_{overt})$ には Sasano et al. (2008) の確率的格解析モデルから得られる表層の係り受けの確率を用い， ϕ_{case} を構成する各素性ベクトルには Hangyo et al. (2013) の素性を利用した．

3.1.2 前処理

先行研究 (Sasano and Kurohashi, 2011; Hangyo et al., 2013) と同様に，まず文書全体に対して形態素解析，固有表現抽出，構文解析を行う．これには JUMAN Ver.7.01[3]，KNP Ver.4.16[4]，CaboCha Ver.0.69[5]を用いた．その後，文頭から出現順に述語単位でゼロ照応解析を行う． 3.1.1 で示したように，各述語項構造候補は格フレーム cf とその格フレームの格スロットと照応先の対応付け a として素性ベクトルを用いて表現する．

3.2 語の分散表現の導入

本研究では，上記のベースラインモデルで用いる素性に加えて，語の分散表現を用いた素性を組み込む．具体的には，格フレーム中の格要素の例から，その述語が格要素としてとりうる語の分散表現を計算し，これを新しい素性として導入する (山城ら, 2017)．格要素の分散表現を計算する手法として使用したのは「格フレーム内平均ベクトル」,「述語内平均ベクトル」の二種類である．語の分散表現を生成するモデルとしては word2vec (Mikolov et al., 2013) を使用した[6]．

「格フレーム内平均ベクトル」とは京大格フレームの情報を用いて，格フレームの格 c の格要素となりうる語 (w) から word2vec を用いてそのベクトル表現 (ϕ_w) を計算し，これらの語ベクトル (ϕ_w) を w が格 c をともなって出現する回数 ($count(cf, c, w)$) (図 1 の「出現回数」の列) で重み付けした平均ベクトル ($\overline{\phi}_{cf,c}$) である．「述語内平均ベクトル」とは「格フレーム内平均ベクトル」の情報 ($\overline{\phi}_{cf,c}$) と述語に対応する出現回数 ($count(p, c, w)$) の総和の情報から述語 p と格 c ごとに対応する平均ベクトル ($\overline{\phi}_{p,c}$) である．

4 格フレーム中の項の分散表現を用いた候補削減手法

4.1 候補の削減について

文間ゼロ照応に際して，いくつかの先行研究ではそれぞれに候補削減の基準を設定している．Sasano and Kurohashi (2011) は述語が含まれる文より 3 文前までに出現する格要素の先行詞候補をすべて含めている．一方で Imamura et al. (2009) は 1 文前までに現れる述語の項として選ばれた項のみを対象としている．この他に Ouchi et al. (2015) は述語項構造解析において，同文中の複数述語と複数名詞の組合せとして考えうるすべて二部グラフのうち，その局所解を山登り法で探索している．

我々は Ouchi et al. (2015) の山登り法を参考に，3.2 で導入した二種類の平均ベクトルを使用し，格フレームと項の組合せ候補を効率的に削減する手法を提案する．この候補削減の目的は計算時間の抑制と正例・負例のデータ数の非対称性の解消である．正例から遠い多数の負例は学習の精度を悪化させると考えられるので，これをできる限り削減することを狙っている．

4.2 格フレーム中の平均ベクトルを用いた山登り法による候補削減

Sasano and Kurohashi (2011) の基準を参考に，ゼロ代名詞となる格要素の先行詞候補は述語が含まれ

[3] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[4] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
[5] https://taku910.github.io/cabocha/

[6] 日本語 wikipedia (2016-09-20) の本文全文から取得した約 100 万記事に対して，次元数を 500，window を 15 として学習させることで得られたモデルを使用した．

アルゴリズム 1 候補削減アルゴリズム

Input:
the predicate p to be analyzed,
the set of case frames F_p coresponding to p,
the set of cases C,
the set of nouns E appearing up to the previous three sentences.
Output:
optimal cf_p, e_c for the analyzed p and each case $c \in C$.

1: **for** each case $c \in C$ **do**
2: $e_c^{(0)} \Leftarrow \underset{e \in E}{\text{argmax}} \cos(\bar{\phi}_{p,c}, \phi_e)$
3: **end for**
4:
5: $cf^{(0)} \Leftarrow \underset{cf \in F_p}{\text{argmax}} \sum_{c \in C} \text{Pseudo-Score}(cf, e^{(0)})$
6: $t \Leftarrow 0$
7: **repeat**
8: **for** each case $c \in C$ **do**
9: $e_c^{(t+1)} \Leftarrow \underset{e_c \in E}{\text{argmax}}\ \text{Pseudo-Score}(cf^{(t)}, e_c)$
10: **end for**
11:
12: $cf^{(t+1)} \Leftarrow \underset{cf \in F_p}{\text{argmax}} \sum_{c \in C} \text{Pseudo-Score}(cf, e_c^{(t+1)})$
13: $t \Leftarrow t + 1$
14: **until** $e_c^{(t)} = e_c^{(t+1)}$ and $cf^{(t)} = cf^{(t+1)}$
15: **return** $\tilde{cf} \Leftarrow cf^{(t)}, \tilde{e_c} \Leftarrow e_c^{(t)}$ for each case $c \in C$
16:
17: **function** PSEUDO-SCORE(cf, e)
18: $score \Leftarrow 0$
19: **for** each case $c \in C$ **do**
20: $score \Leftarrow score + P(p, cf, e, c)$
21: $score \Leftarrow score + \cos(\bar{\phi}_{cf,c}, \phi_e)$
22: $score \Leftarrow score + 0.5 \times d_{p,e}$
23: **end for**
24:
25: **return** *score*
26: **end function**

る文より 3 文前までのみを範囲として候補削減を行っている．これより以前に含まれる名詞についてはそもそも候補削減の対象としておらず，ゼロ代名詞の候補となることもない．この範囲設定は計算量の抑制を目的とする格要素の分布の調査に基づいた制限である (Sasano and Kurohashi, 2011; Hangyo et al., 2013)．

上記の基準で収集した格要素の先行詞候補に対して，さらに平均ベクトルを用いた候補の削減を行う．アルゴリズム 1 にその概要を示す (ただし，$d_{p,e}$ は述語 p と項候補 e の間の文数である)．

ある一つの述語にはその語義の曖昧性を反映した複数の格フレームが存在する．それぞれの格フレームに対応する格フレーム内平均ベクトルはその格フレームの選択選好を反映しているため，これと今見ている項候補ベクトルの距離が近いほど，その項候補は対象格フレームの格スロットにより埋まりやすいと言える．このアルゴリズムは与えられた述語に対して，二者間の距離が最も近くなる格フレームと項候補の組合せを探索している．

まず初期値として各格に埋まりうる項を仮に定める．これには 3.2 で示した述語内平均ベクトルを用いて，格要素候補の分散表現とのコサイン距離を求め，これが最小となる，対象述語に埋まる項群に最も近い項を初期項とする (行 1-3)．次にこれらの初期項に対して最適な格フレームを格フレームの初期値とする (行 5)．ここでスコアとして用いるのは 3.2 で示した格フレーム内平均ベクトルとの単純なコサイン距離のみではなく，これに加えて，京大格フレームに基づく (述語，格フレーム，深層格，項) の組合せの出現確率と，項と述語の間の文数を考慮に含めている (行 17-26)．このスコアの係数は経験的に定めた．以降，格フレームを固定して項を探索するフェーズと項を固定して格フレームを探索するフェーズを繰り返し，格フレームと項が更新されなくなればループを抜ける (行 6-15)．このアルゴリズムでは返り値として最もスコアの高い格フレームと項の組合せを返すが，実際にはループ中の毎回の探索過程で計算した項候補のうち 3 ベストまでを候補として保存し，他は保存しないようにしている．最終的な出力は探索の過程で保存されたすべての格フレームと項の組合せである．この提案手法により，正解を候補に残しつつ，約 1000 分の 1 まで解析候補を削減することに成功した．

5 評価実験

5.1 データ

実験データとして，BCCWJ (Maekawa et al., 2014) のコアデータ[7]と NAIST Text Corpus (NTC) (Iida et al., 2007) を使用した．BCCWJ は，13 ジャンルにまたがって構築された約一億語からなる日本語均衡コーパスである．このうちの約 100 分の 1 にあたる約 2,000 文書のコアデータに対しては，人手による述語項構造と照応関係の付与が行われており，これは新聞，雑誌，書籍，白書，Yahoo!知恵袋，Yahoo!ブログの 6 ジャンルにまたがっている．これをジャンルの偏りがないよう約 5 分の 4 を訓練データ，残りを評価用データとして使用した，NTC のデータ分割については (Taira et al., 2008; Imamura et al., 2009) と同じ分割方法を採択した．

ただし，対象とした述語は動詞のみで，形容詞，事態性名詞は扱っていない．また，格が交代する受身，使役などの助動詞を伴って現れる述語も今回は対象としていない．

5.2 結果と考察

BCCWJ での実験結果を表 2 に，NTC での実験結果を表 3 に示す．

提案手法についてのみ比較した時，全体的に NTC における結果がやや劣る．これは京大格フレームの作成に利用された Web コーパスと新聞コーパスに出現する語句の違いによるものだと考えられる．また，NTC における結果については Sasano and Kurohashi (2011)， Matsubayashi and Inui (2017) のモデルとそれぞれ比較した．ただし，我々が文内文間の動詞のみを対象としているのに対し， Sasano and Kurohashi (2011) は文内文間の動詞，形容詞を取り扱っており，

[7] http://pj.ninjal.ac.jp/corpus_center/bccwj/

表 2: BCCWJ を用いた実験結果の精度 (F 値)

	文内				文間				All			
格	ガ格	ヲ格	ニ格	All	ガ格	ヲ格	ニ格	All	ガ格	ヲ格	ニ格	All
事例数	3,137	1,458	873	5,468	2,359	495	243	3,097	5,496	1,953	1,116	8,565
提案手法	.535	.717	.741	.614	.128	.105	.131	.124	.391	.586	.627	.468

表 3: NTC を用いた実験結果の精度 (F 値)

	文内				文間				All			
格	ガ格	ヲ格	ニ格	All	ガ格	ヲ格	ニ格	All	ガ格	ヲ格	ニ格	All
事例数	11,559	7,472	4,389	23,420	2,810	229	142	3,181	14,369	7,701	4,531	26,601
提案手法	.227	**.271**	**.120**	.224	.071	.020	.014	.058	.193	.243	.111	.196
(Sasano and Kurohashi, 2011)	**.395**	.175	.089		**.244**	**.066**	**.026**					
(Matsubayashi and Inui, 2017)	.565	.447	.160	.537								

Matsubayashi and Inui (2017) は直接係り受け関係と文内の動詞，形容詞，事態性名詞を取り扱っているため単純な比較はできない．表 3 中の事例数は動詞のみのものである． Sasano and Kurohashi (2011) と比較した時，文内のヲ格ニ格については勝っているものの，ガ格で劣っている．これは新聞コーパスのガ格に出現する語の種類が比較的限定され，位置関係からも推測しやすいため， Sasano and Kurohashi (2011) が訓練に使用した Web コーパスが新聞に似た特徴を持っていた場合，そのドメインの偏りが有利に働いた可能性がある．しかし全体としては概ね同程度の性能を示していると言える．Matsubayashi and Inui (2017) と比較した時，文内ゼロ照応においては述語の含まれる同文中のみを探索すればよいのに対して，問題を文間にまで拡張した時の項候補は前文中すべてを探索する必要がある．その項候補の増大を考えると， Matsubayashi and Inui (2017) と比べて，全体的に約半分以下の性能しか示していないが，提案した項候補削減手法が上手く機能していると考えられる．

6 おわりに

本稿では分散表現で平均化した格フレームを用いた文内，文間ゼロ照応解析の項候補削減手法を提案，大規模な多ドメインコーパスによる訓練を可能にし，提案したモデルが既存研究と同程度の性能を達成していることを確認した．

謝辞

(Hangyo et al., 2013) に関して詳細な情報をご教示くださった萩行正嗣氏，(Matsubayashi and Inui, 2017) との比較実験にご助力くださった松林優一郎氏，(Ouchi et al., 2017) の全体像についてご教示くださった大内啓樹氏に厚く御礼申し上げます．

参考文献

Hangyo, M., Kawahara, D., and Kurohashi, S. Japanese Zero Reference Resolution Considering Exophora and Author/Reader Mentions.. Proceedings of EMNLP, pp. 924-934, 2013.

Iida, R., Komachi, M., Inui, K., and Matsumoto, Y. Annotating a Japanese text corpus with predicate-argument and coreference relations. Proceedings of the Linguistic Annotation Workshop, pp. 132-139, 2007.

Imamura, K., Saito, K., and Izumi, T. Discriminative approach to predicate-argument structure analysis with zero-anaphora resolution. ACL-IJCNLP, pp. 85-88, 2009.

Joachims, T. Training linear SVMs in linear time. Proceedings of the 12th ACM SIGKDD, 2006.

Kawahara, D. and Kurohashi, S. A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis. ACL, pp. 176-183, 2006.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y. Balanced corpus of contemporary written Japanese. Language Resources and Evaluation, Vol. 48, No. 2 , pp. 345-371, 2014.

Matsubayashi, Y. and Inui, K. Revisiting the Design Issues of Local Models for Japanese Predicate-Argument Structure Analysis. IJCNLP, pp. 128-133, 2017.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

Ouchi, H., Shindo, H., Duh, K., and Matsumoto, Y. Joint Case Argument Identification for Japanese Predicate Argument Structure Analysis.. ACL-IJCNLP, pp. 961-970, 2015.

Ouchi, H., Shindo, H., and Matsumoto, Y. Neural Modeling of Multi-Predicate Interactions for Japanese Predicate Argument Structure Analysis. ACL, pp. 1591-1600, 2017.

Sasano, R. and Kurohashi, S. A Discriminative Approach to Japanese Zero Anaphora Resolution with Large-scale Lexicalized Case Frames.. IJCNLP, pp. 758-766, 2011.

Sasano, R., Kawahara, D., and Kurohashi, S. A fully-lexicalized probabilistic model for Japanese zero anaphora resolution. COLING, pp. 769-776, 2008.

Shibata, T., Kawahara, D., and Kurohashi, S. Neural Network-Based Model for Japanese Predicate Argument Structure Analysis.. ACL, p. 12351244, 2016.

Taira, H., Fujita, S., and Nagata, M. A Japanese predicate argument structure analysis using decision lists. EMNLP, pp. 523-532, 2008.

山城颯太, 西川仁, 徳永健伸. 分散表現による格フレームの格要素の汎化を利用したゼロ照応解析. 言語処理学会第 23 回年次大会発表論文集, pp.206-209, 2017.