Entity-Centric な述語項構造解析・共参照解析の同時学習

柴田 知秀　　黒橋 禎夫
京都大学　科学技術振興機構 CREST
{shibata, kuro}@i.kyoto-u.ac.jp

1　はじめに

述語項構造解析は情報抽出や質問応答など様々なアプリケーションにとって重要となる基礎解析である。述語項構造解析には様々な手がかり・知識が必要となり形態素解析や構文解析などに比べると難しいタスクではあるが、ニューラルネットワークを用いることにより精度が向上してきている [5, 2, 3] 。

述語項構造解析は項が文内と文間にある場合に大別することができる。文間の解析は非常に難しいことから、近年は項が文内にある場合のみを対象にしている。しかし、アプリケーションでの利用などを考えると当然ながら文間の解析も扱えることが望ましい。文間の解析の場合、文内の解析で有効である述語と項の間のパスなどの情報が使えないため、解析が難しくなる。文間の解析の精度を向上させるためには談話のモデリング、つまり文章中で何が話題の中心であるかなどを捉える必要がある。

述語項構造解析と同様のレイヤの解析に共参照解析がある。共参照解析では照応詞と先行詞候補の間でスコアを計算し、最もスコアの高いものを先行詞とすることで解析が行われる。その際に、これまでの解析から得られるクラスタ (entity) の情報が有効となる [6, 1]。entity の表現方法はそれほど自明ではないが、Wiseman らは RNN で entity の embedding を計算し、これを用いることで精度向上を達成している [6]。

本論文では entity-centric な述語項構造解析・共参照解析の同時学習手法を提案する。共参照解析で利用されている entity の情報を述語項構造解析でも利用する。そして、両解析において、ある entity を指した場合にその entity の embedding を更新し、情報を蓄積する。例えば以下の文章を考える。

(1)　コワリョフ氏$_1$ は正式な党員ではないが、一九九二年の選挙でロシア共産党から (ϕ_1 ガ) 立候補し (ϕ_1 ガ) 当選した。同氏$_{=1}$ は当選までロシア内務省法律研究所の教授を務めていた法律学者。九四年に大統領令でチェチェン問題に関する人権監視委員会の委員長に (ϕ_1 ガ) 任命された。チェチェンへの軍事介入に関しては、... エリツィン大統領の立場を (ϕ_1 ガ) 支持していた。

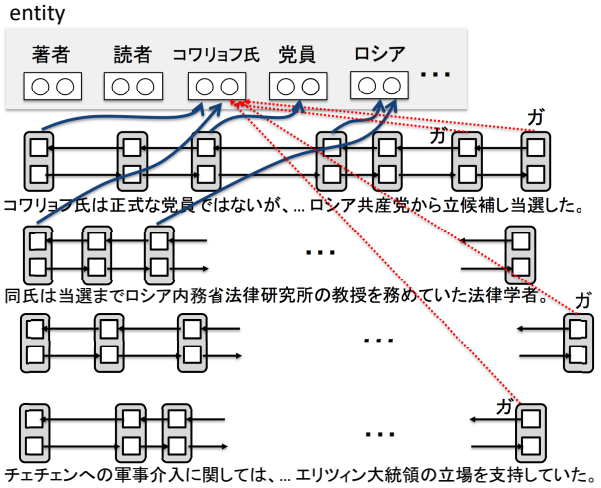

図 1: 提案手法の概要

最終文の「支持した」のガ格の解析において、1 文目の「コワリョフ氏」(またはそれと共参照関係にある 2 文目の「同氏」) を指す必要がある。しかし、文間の解析において数ある候補から正しい先行詞を同定するのは難しい。しかし、「コワリョフ氏」は 2 文目の「同氏」と共参照の関係にあることや、1 文目の「立候補」や「当選」のガ格、3 文目の「任命された」のガ格とゼロ照応の関係にあることから話題の中心であることがわかる。本研究ではそれらの情報を「コワリョフ氏」の entity の embedding として蓄積しておき、その上で最終文の「支持した」の解析を行うことで、述語項構造解析の精度向上を図る。

提案手法の概要を図 1 に示す。まず、入力文書を encoding する。次に、文章の先頭から共参照解析ならびに述語項構造解析を行う。各 entity には embedding を与えておき、両解析時にそれを考慮する。そして、両解析で指された場合に embedding を更新し、情報

を蓄積していく。

実験を行ったところ、提案手法は述語項構造解析・共参照解析の精度を向上させ、特に文間ゼロ照応解析の精度を大きく向上することができた。

2 ベースモデル

まず、入力文書に対して encoding し、その後、文章の先頭から用言の場合は述語項構造解析を、それ以外の場合は共参照解析を行う。本節ではまずベースとなるモデルについて述べる。

2.1 入力文書の Encoding

共参照解析・述語項構造解析は基本句 [1] を単位に行われるので、まず基本句の embedding を計算する。基本句内の各単語を単語 (表記)、品詞、品詞細分類、活用型の embedding を連結したもので表す。そして、基本句の単語列の embedding に対して CNN を適用することで基本句の embedding を得る [2](i 番目の基本句の embedding を $\boldsymbol{x_i}$ と表す)。

そして、各文において、文脈を考慮した embedding を得るために、基本句の embedding を bi-direcitional LSTM で読む。

$$\begin{aligned}\overrightarrow{\boldsymbol{h}}_i &= \overrightarrow{LSTM}(\boldsymbol{x}_i, \overrightarrow{\boldsymbol{h}}_{i-1}) \\ \overleftarrow{\boldsymbol{h}}_i &= \overleftarrow{LSTM}(\boldsymbol{x}_i, \overleftarrow{\boldsymbol{h}}_{i+1})\end{aligned} \tag{1}$$

i 番目の基本句の embedding は順方向と逆方向を連結したものとする ($\boldsymbol{h}_i = [\overrightarrow{\boldsymbol{h}}_i; \overleftarrow{\boldsymbol{h}}_i]$)。

2.2 共参照解析

照応詞に対して先行詞候補の中から最もスコアの高いものを先行詞として採用する mention-ranking model で共参照解析を行う。照応詞 m が先行詞候補 a を指す良さを表すスコア $s_{coref}(a, m)$ を計算する。先行詞候補は m より前の mention、外界照応、NA(先行詞なし) であり、$s_{coref}(a_j, m_i)$ は以下の式で計算される。

$$s_{coref}(a_j, m_i) = W_2^{coref} ReLU(W_1^{coref} \boldsymbol{v}_{input}^{coref}) \tag{2}$$

ここで、$\boldsymbol{v}_{input}^{coref}$ は共参照解析のための入力を表し、以下を連結したベクトルとする。

- 先行詞と照応詞それぞれに対応する embedding
- 先行詞と照応詞の文字列が完全一致または部分一致しているかどうか
- 先行詞と照応詞が何文離れているか (0,1,2,3+文)
- KNP の同義語辞書にマッチするかどうか

先行詞が NA の場合は上記から先行詞ならびに先行詞・照応詞ペアから得られる情報を除き、パラメータの異なるネットワークでスコアを計算する。

ロス関数は以下の max-margin で与えられる。

$$\mathcal{L}_{coref} = \sum_i^{N_m} \max_{a_j \in \mathcal{A}(m_i)} (1 + s_{coref}(a_j, m_i) - s_{coref}(\hat{t}_i, m_i)) \tag{3}$$

ここで、N_m は文書中の mention の数、$\mathcal{A}(m_i)$ を mention m_i の先行詞候補、$\hat{t}_i$ は正解先行詞の中で最も score の高いものを表す。

2.3 述語項構造解析

述語項構造解析は格解析 (述語と係り受け関係にある項のうち明示されていない格を同定する解析) とゼロ照応解析 (ゼロ代名詞を認識しその先行詞を同定する処理) からなる。述語においてどちらの解析も統一的に、ある格に対して全ての候補の中から項となるものを一つ選ぶと考える。候補は文章中に現れている基本句、著者や読者などの外界照応、対応する項がないことを示す NULL からなる。

述語 p_i の格 c が項 a_j となる確率を以下の式で与える。

$$P(c = a_j | p_i) = \frac{\exp(s_{PAS}(a_j, p_i, c))}{\sum_k \exp(s_{PAS}(a_k, p_i, c))} \tag{4}$$

$s_{PAS}(a_j, p_i, c)$ は以下の式で計算される。

$$s_{PAS}(a_j, p_i, c) = W_2^{PAS} \tanh(W_1^{PAS} \boldsymbol{v}_{input}^{PAS}) \tag{5}$$

ここで、$\boldsymbol{v}_{input}^{PAS}$ は述語項構造解析のための入力を表し、以下を連結したベクトルとする。

- 述語 p_i ならびに項 a_j の embedding
- path embedding[4]: 候補となる項から述語に向かって構文木上に LSTM で読み、最後のステップの隠れ層を path embedding とする [3]
- 選択選好: 述語項構造解析における重要な手がかりの一つに選択選好がある。Shibata らの手法 [5] で選択選好のスコアを計算する
- 述語と項が何文離れているか (0,1,2,3+文)

ロス関数は softmax cross entropy で与えられる [4]。

$$\mathcal{L}_{PAS} = -\sum_i^{N_p} \sum_c \log P(c = \hat{a} | p_i) \tag{6}$$

ここで、N_p は文書中の述語の数、$\hat{a}$ は正解の項を表す。

[1] 基本句は 1 つの自立語と 0 個以上の付属語からなる単位である。

[2] フィルタサイズは 1,2,3 とし、基本句の先頭、末尾である情報を得るために、先頭と末尾には特別な embedding を付与する。

[3] 文間や外界照応の項の場合はゼロベクトルとする。

[4] 述語と直接係り受け関係にあり、その項の格が明示されている場合は解析対象から除く。

3 Entity Embedding

Wiseman らの手法 [6] と同じように、各 entity に embedding を割り当てる。解析で参照されるたびに entity の embedding を更新し、情報を蓄積していく。また解析時には entity の embedding を入力に用いることで、各 entity の情報を参照する。

3.1 Entity Embedding の更新

時刻 i での解析 (entity の embedding の更新を含む) が終わった時点での entity k の embedding を $\boldsymbol{e}_i^{(k)}$ と表記する。共参照解析で entity k が指された場合、$\boldsymbol{e}_i^{(k)}$ を以下のように更新する。

$$\boldsymbol{e}_i^{(k)} \leftarrow LSTM_e(\boldsymbol{h}_i, \boldsymbol{e}_{i-1}^{(k)}) \tag{7}$$

図 1 の例では、2 文目の「同氏」が 1 文目の「コワリョフ氏」と共参照の関係にある場合、「コワリョフ氏」の entity の embedding を上記の式で更新する。

これに加えて本研究ではゼロ照応で指された場合も embedding を更新する。ゼロ照応の場合、照応詞がないので、述語 $\boldsymbol{h}_i$ が格 c で entity k を指した場合、以下のように更新する。

$$\boldsymbol{e}_i^{(k)} \leftarrow LSTM_e(W_c\boldsymbol{h}_i, \boldsymbol{e}_{i-1}^{(k)}) \tag{8}$$

図 1 の例では、1 文目の「立候補」のガ格が「コワリョフ氏」である場合、「コワリョフ氏」の entity の embedding を更新する。共参照解析・述語項構造解析ともに、参照されなかった entity の embedding は更新されない ($\boldsymbol{e}_i^{(l)} \leftarrow \boldsymbol{e}_{i-1}^{(l)} (l \neq k)$)。

3.2 Entity Embedding を考慮した解析

2.2 節、2.3 節で説明した共参照解析、述語項構造解析において entity の embedding を考慮する。

共参照解析においては Wiseman らの手法と同様に entity を考慮したスコア $score_{coref}^{entity}$ を計算し、2.2 節でのスコア $score_{coref}$ に加算する。mention m_j が entity k に属することを $z_j = k$ と表すことにすると、このスコアは次式で計算される。

$$score_{coref}^{entity}(a_j, m_i) = \begin{cases} \boldsymbol{h}_i^{\mathrm{T}} \boldsymbol{e}_{i-1}^{(z_j)} & (a_j \neq \epsilon) \\ NA(m_i) & (a_j = \epsilon) \end{cases} \tag{9}$$

$$NA(m_i) = \boldsymbol{q}^{\mathrm{T}} \tanh(W_{NA}\begin{bmatrix} \boldsymbol{h}_i \\ \sum_k \boldsymbol{e}_{i-1}^{(k)} \end{bmatrix}) \tag{10}$$

述語項構造解析においては項 a に対応する entity の embedding を入力層に加える。

学習時は共参照解析・述語項構造解析ともに正解を用いて entity の embedding を更新し、テスト時はシステムの出力で entity の embedding を更新する。

4 実験

4.1 実験設定

評価は京都大学ウェブ文書リードコーパス (**ウェブ**: 約 5,000 文書, 15,000 文) と京都大学テキストコーパス (**新聞**: 約 550 文書, 5,000 文[5]) の 2 種類のコーパスで行った。対象の格はガ、ヲ、ニ、ガ 2 格の 4 種類とし、外界照応として著者、読者、不特定:人を考えた。評価は F 値とし、述語項構造解析・共参照解析ともに共参照のリンクを用いて評価をゆるめた。また、述語項構造解析において著者と不特定:人の区別は非常に難しいので、これらの違いは正解とみなした。

実装詳細

エポック数は 10 とした。checkpoint ensemble を採用し、dev セットにおいて述語項構造解析と共参照解析解析の F 値の和のベスト 5 のモデルのパラメータを平均し、テスト時に用いた。

単語ベクトルの次元は 100 で、日本語 1 億ページで学習した embedding を初期値にした。品詞、品詞細分類、活用型の次元はそれぞれ 10 とし、ランダムに初期化した。すべてのニューラルネットワークにおいて隠れ層の次元を 100 とした。最適化手法として Adam を用いた。実験結果は 5 回の試行の平均とした。

4.2 実験結果

次の 3 つの手法を比較した。「ベースライン」は 2 節で説明したモデルで、「+entity (coref)」は共参照解析において entity の embedding を利用するもので、Wiseman らの手法の有効性をみるためのものである。「+entity (PAS)」は提案手法であり、ゼロ照応で指された場合にも entity の embedding を更新し、述語項構造解析で entity の embedding を利用するものである。

格解析・ゼロ照応解析・共参照解析の精度を表 1 に示す。格解析の精度はどの手法もほぼ同じである。「+entity (coref)」で共参照解析の精度が向上しており、提案手法である「+entity (PAS)」でゼロ照応解析・共参照解析ともに精度が向上していることがわかる。また、格解析・ゼロ照応解析の格別ならびにゼロ照応解析の先行詞の位置別の精度を表 2 に示す。提案手法では特に文間の精度が大きく向上しており、entity の embedding の有効性を示すことができた。

以下に解析例を示す。

(2) ムルジ$_1$ は、バンダが政府を去った後、ＵＤＦを率いる とともに民主主義の代弁者となった。しかし、ムルジの大統領としての日々は、... 費や

[5] 省略、共参照の関係が付与されている部分のみを利用した。

手法	格解析	ゼロ照応解析	共参照解析	格解析	ゼロ照応解析	共参照解析
	ウェブ			新聞		
ベースライン	0.881	0.523	0.678	0.895	0.280	0.546
+entity (coref)	**0.885**	0.519	0.689	**0.897**	0.296	0.555
+entity (PAS)	0.881	**0.561**	**0.691**	0.894	**0.355**	**0.565**

表 1: 格解析・ゼロ照応解析・共参照解析の精度 (値は F 値を表す)

格	手法	格解析	ゼロ照応解析				格解析	ゼロ照応解析			
			すべて	文内	文間	外界		すべて	文内	文間	外界
		ウェブ					新聞				
ガ	ベースライン	0.941	0.595	0.401	0.054	**0.729**	0.943	0.319	0.443	0.038	0.337
	+entity (coref)	**0.943**	0.589	0.409	0.057	0.723	0.943	0.342	0.454	0.034	0.398
	+entity (PAS)	0.937	**0.631**	**0.453**	**0.408**	**0.729**	**0.944**	**0.392**	**0.473**	**0.224**	**0.493**
		(1,433)	(2,005)	(333)	(394)	(1,278)	(893)	(1,011)	(443)	(391)	(177)
ヲ	ベースライン	0.847	0.186	0.284	0.058	0.000	0.648	0.016	0.030	0.000	0.000
	+entity (coref)	**0.849**	0.205	**0.298**	0.074	0.000	0.652	0.016	0.029	0.000	0.000
	+entity (PAS)	0.841	**0.232**	0.290	**0.182**	0.000	**0.654**	**0.046**	**0.060**	**0.032**	0.000
		(285)	(214)	(96)	(105)	(13)	(105)	(96)	(40)	(56)	(0)
ニ	ベースライン	0.325	0.393	0.067	0.023	0.526	0.260	0.185	**0.026**	0.000	0.382
	+entity (coref)	0.369	0.393	0.098	0.013	0.528	**0.290**	0.179	0.008	0.000	0.376
	+entity (PAS)	**0.399**	**0.436**	**0.121**	**0.121**	**0.561**	0.243	**0.307**	0.008	**0.007**	**0.541**
		(99)	(570)	(80)	(149)	(341)	(26)	(286)	(80)	(91)	(115)
ガ 2	ベースライン	0.385	0.188	0.199	0.000	0.232	0.150	0.000	0.000	0.000	0.000
	+entity (coref)	0.442	0.188	**0.210**	0.000	0.227	0.126	0.000	0.000	0.000	0.000
	+entity (PAS)	**0.496**	**0.291**	0.196	**0.122**	**0.360**	**0.234**	0.000	0.000	0.000	0.000
		(107)	(139)	(28)	(28)	(83)	(13)	(35)	(15)	(13)	(7)
すべて	ベースライン	0.881	0.523	0.341	0.046	0.673	0.895	0.280	0.383	0.028	0.348
	+entity (coref)	**0.885**	0.519	0.351	0.049	0.669	**0.897**	0.296	0.392	0.025	0.386
	+entity (PAS)	0.881	**0.561**	**0.379**	**0.318**	**0.680**	0.894	**0.355**	**0.400**	**0.183**	**0.507**
		(1,924)	(2,928)	(537)	(676)	(1,715)	(1,037)	(1,428)	(578)	(551)	(299)

表 2: 格解析・ゼロ照応解析の格別ならびにゼロ照応解析の先行詞の位置別の精度 (括弧内の数字は項の数を表す)

> された日々でもあった。特に、国中に飢饉をもたらした旱魃が始まる直前に、他国へトウモロコシの備蓄を 売却 していたことが問題となった。

1 文目の「率いる」のガ格はベースライン、提案手法ともに「ムルジ」と解析できていたが、ベースラインでは最終文の「売却」のガ格を誤って「不特定:人」と解析してしまった。提案手法では「ムルジ」が共参照・ゼロ照応で指されていることを捉えることができ、正しく「ムルジ」と解析することができた。

5 おわりに

本論文では entity-centric な述語項構造解析・共参照解析の同時学習手法を提案した。提案手法は両解析の精度を向上させ、特に文間ゼロ照応解析の精度を大きく向上することができた。今後の課題として橋渡し参照などを行うことや事態間関係知識などの外部知識を取り入れることなどがあげられる。

謝辞

本研究は科学技術振興機構 CREST「知識に基づく構造的言語処理の確立と知識インフラの構築」の支援のもとで行われた。

参考文献

[1] Kevin Clark and Christopher D. Manning. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 643–653, 2016.

[2] Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, Canasai Kruengkrai, and Julien Kloetzer. Intra-sentential subject zero anaphora resolution using multi-column convolutional neural network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1244–1254, 2016.

[3] Hiroki Ouchi, Hiroyuki Shindo, and Yuji Matsumoto. Neural modeling of multi-predicate interactions for Japanese predicate argument structure analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1591–1600, 2017.

[4] Michael Roth and Mirella Lapata. Neural semantic role labeling with dependency path embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1192–1202, 2016.

[5] Tomohide Shibata, Daisuke Kawahara, and Sadao Kurohashi. Neural network-based model for Japanese predicate argument structure analysis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1235–1244, 2016.

[6] Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. Learning global features for coreference resolution. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 994–1004, 2016.