

B4IM2016

修士論文

項の先行出現文脈を考慮した 分散表現に基づく選択選好モデル

大野 雅之

2016年3月25日

東北大学 大学院
情報科学研究科 システム情報科学専攻

本論文は東北大学 大学院情報科学研究科 システム情報科学専攻に
修士(工学) 授与の要件として提出した修士論文である。

大野 雅之

審査委員：

乾 健太郎 教授 (主指導教員)

周 暁 教授

大町 真一郎 教授

岡崎 直観 准教授 (副指導教員)

項の先行出現文脈を考慮した 分散表現に基づく選択選好モデル*

大野 雅之

内容梗概

述語の選択選好性のモデル化は、述語項構造解析・省略解析を始めとした意味解析において重要な基盤技術の一つである。これまでの研究では、「述語の選択選好性は項となる名詞の意味的な性質にのみ依拠する」という仮定して選好性を学習してきた (e.g., man は arrest の目的語になりうる)。しかしながら、省略解析などの談話解析への応用を考えると、談話内での名詞の言及のされ方まで含めて選好性を計算できることが望ましい (e.g., 悪事を犯した man は arrest の目的語になりうるが、善良な man はなりにくい)。

本研究では、ニューラルネットワークに基づく選択選好モデル [Van de Cruys 2014] を拡張し、名詞の意味的な性質に加え、談話内での言及のされ方を分散表現で表現することにより、名詞の出現文脈を考慮した述語の選択選好モデルを提案した。評価実験では、代名詞照応解析への応用を見据え、代名詞に対する先行詞候補のランキング問題に基づく評価を行い、項の先行出現文脈を用いることの有効性を確認した。

キーワード

自然言語処理, 談話解析, 照応・省略解析, 文脈理解, 選択選好, 分散表現

*東北大学 大学院情報科学研究科 システム情報科学専攻 修士論文, B4IM2016, 2016年3月25日.

目次

| | | |
|----------|-----------------------|-----------|
| 1 | はじめに | 1 |
| 2 | 関連研究 | 4 |
| 2.1 | 述語の選択選好性に関する研究 | 4 |
| 2.1.1 | シソーラスによるクラスタリング | 4 |
| 2.1.2 | 確率的潜在変数モデル | 4 |
| 2.1.3 | 分散表現に基づくモデル | 5 |
| 2.1.4 | 複数の項の依存関係を考慮したモデル | 6 |
| 2.2 | 事象間関係知識を用いた談話解析に関する研究 | 6 |
| 3 | 文脈を考慮した選択選好モデル | 8 |
| 3.1 | ニューラルネットワークを用いたモデル | 8 |
| 3.2 | 項の先行文脈を考慮した選択選好モデルの構築 | 10 |
| 3.3 | 定式化 | 11 |
| 3.4 | 訓練事例の作成 | 13 |
| 3.5 | モデルパラメータの学習 | 17 |
| 4 | 評価実験 | 19 |
| 4.1 | 実験設定 | 19 |
| 4.1.1 | 評価タスク | 19 |
| 4.1.2 | 評価データの作成 | 20 |
| 4.1.3 | 比較対象 | 22 |
| 4.2 | 実験結果 | 22 |
| 4.2.1 | 先行文脈の考慮による性能変化 | 22 |
| 4.2.2 | 訓練データ量と性能 | 24 |
| 4.2.3 | 単語ベクトルと性能 | 26 |
| 4.3 | 事例分析 | 28 |
| 5 | おわりに | 29 |

目次

| | | |
|---|--------------------------------------|----|
| 1 | SVO モデルのネットワーク構造 | 9 |
| 2 | 提案手法の概要 | 12 |
| 3 | Stanford CoreNLP による解析結果 | 15 |
| 4 | 訓練に用いる共参照事例数の変化による MRR の変化 | 25 |
| 5 | 訓練に用いる共参照事例数の変化による MQ の変化 | 25 |
| 6 | 訓練に用いる共参照事例数の変化による ACC の変化 | 26 |

表目次

| | | |
|---|--|----|
| 1 | 本研究で扱う先行文脈 | 10 |
| 2 | 動詞・名詞・形容詞として用いた品詞タグ | 16 |
| 3 | 逆接や否定を表す係り受け | 17 |
| 4 | PropBank による意味役割と主語, 述語, 目的語との対応 | 21 |
| 5 | 先行詞候補ランキング問題での MRR | 23 |
| 6 | 先行文脈を構成する述語ごとの学習事例数 | 23 |
| 7 | 単語ベクトル及び隠れ層の次元数を変化させた際の結果 | 27 |

1 はじめに

自然言語処理は、日本語や英語といった自然言語を計算機によって処理し活用する分野であり、web上に存在する大規模なテキストデータを処理する技術として注目を集めている。計算機で言語を正しく処理するためには、述語項構造解析や省略解析など様々な意味解析技術が必要であり、これらに関する研究が盛んに行われている。これらの解析には解析対象となる文から得られる統語的情報に加えて、様々な語彙知識が用いられる。その中に、**述語の選択選好性**の知識がある。

述語の選択選好性とは、ある述語の各格がどのような名詞句を取りやすいかといった選好性のことである。例えば、「食べる」という述語の目的格は、「りんご」のように食べ物を意味する名詞句を取ることがほとんどであり、「時計」のように食べ物ではない名詞句を取るとは極めて稀である。このように、述語の各格が取りうる名詞句には意味的性質による偏りが存在しており、述語の選択選好性の知識は述語の性質を表す重要な知識となっている。自然言語処理の分野において、述語の選択選好性の知識は述語項構造解析や省略解析を始めとした意味解析に有効な情報とされ、これを利用した多くの研究がなされてきた [1, 2, 3, 4, 5]。

述語の選択選好性は述語によってそれぞれ異なるため、人手によりその性質を全てを書き下すのは現実的でない。このため先行研究では、大規模コーパスから、述語と、その述語の格を埋める名詞（以後、**項**と呼ぶ）の用例を獲得し、この共起の統計値から選択選好のモデルを構築するアプローチが採られてきた。また、共起用例をそのまま保持するだけでは、出現頻度の低い述語に関して選好性の知識を正しく得られないという問題が起こる（**データスパースネス問題**）ため、ソーラスの利用 [6] や単語の分散表現化 [7, 8, 9]、潜在変数を用いた確率モデル化 [10, 11, 12, 13] といった方法で、獲得した述語と項の共起用例をいかに一般化するか、という点でさまざまな工夫がなされてきた。

一方、選択選好性が述語項構造解析や照応解析の際に用いられる知識であることを踏まえて現実的な実用上の状況を考えた場合、実際にはそれぞれの名詞には文章上で様々な言及がなされる。選択選好性によって項としての尤もらしさを判定する際には、単に項となる名詞自身の意味的性質だけではなく、その名詞に対して文章上でどのような言明がなされてきたかを踏まえて選択選好性を判定す

る方が自然である。例えば、次のような照応の問題を解いているという状況を考える。

(1) [*John*_(i)] attacked [*Bob*_(j)]. Police arrested *him*_(i).

この例において、先行詞の候補は *John*_(i) と *Bob*_(j) の二種類であると仮定すると、

“attack” された人より “attack” した人のほうが “arrest” される人として尤もらしい

という選好性により *John* が *him* の先行詞であると判断される。しかし、これまでの選択選好モデルでは、述語の選択選好性が名詞自身との関係の性質だけから決まると仮定しており、上のような文脈の選好性は無視されてきた。これは例文 (1) では、

(2) a. Police arrested [*John*]

b. Police arrested [*Bob*]

の二つの尤もらしさを比較していることになるが、文脈なしにはこれらの項としての尤もらしさに大きな差は現れない。一方、例文 (3a), (3b) のように先行文脈により *John* や *Bob* に付加されている情報を加味して選好性を考慮できれば、正しい解を導き出せるはずである。

(3) a. Police arrested [*John, who attacked Bob*]

b. Police arrested [*Bob, whom John attacked*]

実際に照応解析の先行研究では、この種の先行詞の周辺文脈情報を利用することで照応解析の解析性能が向上することが報告されており [14, 5], 談話への応用を考える上で重要な問題だといえる。

こうした背景のもと本研究では、先行文脈における項に対する言及から得られる、項の付加的な意味情報を用いた述語の選択選好性モデルを構築する。また、先に述べたデータスパースネス問題に対応するために、分散表現に基づく選択選好性モデルをベースとして、項の意味的性質を先行文脈から構成的に計算し、述語の選択選好性を計算する手法を提案する。

本論文の構成は以下の通りである。まず2節において、本研究に関連のある、述語の選択選好性のモデル化および先行文脈を用いた談話解析の研究について述べる。次に3節で本研究のベースとなるニューラルネットワークを用いた選択選好モデルについて説明し、分散表現に基づいて項の周辺文脈から項の意味計算を行う選択選好モデルを構築する。そして、4節では、提案モデルが項の先行文脈を考慮して項の意味的性質を適切に計算できていることを確認するため、評価実験を行ない、その結果について考察する。最後に5節にて、本論文のまとめを述べる。

2 関連研究

本節では、関連研究として、選択選好性のモデル化手法に関する既存研究について、概要を紹介する。また本研究と類似した研究として、事象間関係知識を用いた談話解析についても説明し、本研究との違いを明確にする。

2.1 述語の選択選好性に関する研究

本論文の冒頭で述べたように、述語の選択選好性モデルの研究は、大規模コーパスから自動獲得した述語と項の共起用例を、何らかの方法で一般化するアプローチが主流である。本小節では、これまでに行われてきた一般化の手法について特徴をまとめる。

2.1.1 シソーラスによるクラスタリング

語述語と項の共起情報を用いた選択選好モデルの自動構築の歴史において、初期の頃には人手で整備したシソーラスを用いて一般化する試みがなされてきた。シソーラスを用いた手法では、獲得した共起事例に対して項となる名詞をあらかじめ用意したシソーラスの内容に従ってまとめることで、項の汎化を行っている。Resnik [6] は人手で作成した名詞の類義語クラスタ (WordNet [15] の synset) に基づいて項を汎化し、確率モデルを構築した。シソーラスによる汎化を行うことで確率モデルがスムージングされ、性能が向上することを示した。WordNet を用いた一般化手法は類義語によるクラスタリング以外にも、上位下位関係に着目したクラスタリング [16] や、ベイジアンモデルと組合せた手法 [11] など様々なものが提案されている。

2.1.2 確率的潜在変数モデル

人手で作成したシソーラスには対義語のように自動獲得が難しい関係知識も扱えるため有用な資源であるが、言語資源を人手で作成する作業は非常にコストが高い。また、新語が出現や記載漏れがあるたびに資源を更新する必要がある、扱

いが非常に難しい。これを克服するため、人手による資源を与えずに一般化を行う手法が研究されている。

言語資源を用いずに一般化する手法として、潜在変数を用いた手法が広く行われている。潜在変数を用いた手法の一つとして、Rooth ら [10] は最尤推定による確率的潜在変数モデルを提案している。潜在変数を用いることで、出現分布をもとに単語の類似度の計算が可能になり、訓練コーパスに出現しなかった述語と項の組に関する選択選好知識の獲得を可能にしている。確率的潜在変数モデルは他に、Latent Dirichlet Allocation (LDA) を用いた手法 [11, 12] が提案されている。また、日本語を対象とした意味解析において広く使用されている京大格フレームも、潜在変数を用いた手法と類似した方法で、動詞の意味フレームを導出している [17, 13]。

2.1.3 分散表現に基づくモデル

潜在変数モデルにより、外部資源を用いずに単語の類似度を測ることが可能になり、選択選好性モデルの網羅性は格段に向上した。外部資源を用いずに一般化を行う方法は潜在変数モデルに以外にも提案されており、単語を実数値ベクトルで表現する**分散表現**に基づく手法が提案されている [7, 18, 8, 9]。分散表現に基づく手法では、単語を実数値ベクトルで表現することで単語間の類似度の計算が可能になる。また、分散表現に基づく手法は選択選好性モデルに限らず様々な言語処理タスクに用いられ、データスパースネス問題への有効性が報告されており [19, 20]、標準的な対応手段になりつつある。

分散表現に基づく手法では、単語の意味はその単語が出現する文脈によって決まる、という**分布仮説** [21] に則って単語の意味ベクトルを獲得する。ここでの文脈とは、単語の前後数単語や係り受け関係にある単語の情報である。選択選好モデルにおいては、述語項関係を表す係り受け関係から単語ベクトルの獲得を行う。単語ベクトルの構築方法は、以下の2つに大別できる。

- 述語と項の共起行列を次元圧縮 [7, 18, 8]

述語と項の共起情報を埋め込んだ行列（テンソル）を作成し、各単語が対応する行を単語ベクトルとする方法。なお、共起行列は高次元かつ疎な行

列となるため、特異値分解 (Singular Value Decomposition, SVD) 等で次元圧縮して得られた密なベクトルが単語ベクトルとして用いられる。

- ニューラルネットワークを用いた学習 [9]
ニューラルネットワークで選択選好性の識別モデルを学習する過程でパラメータのひとつとして単語ベクトルを学習する方法。

これらの手法は共に高い性能を達成しているが、共起行列から単語ベクトルを構築する手法は共起行列や共起情報を埋め込んだテンソルを事前に用意する必要があり、モデルの拡張性の低さや大量の計算機資源が必要である点で課題が残る。

本研究が構築するモデルも分散表現に基づいたものであり、Van de Cruys が提案したニューラルネットワークを用いた手法 [9] を拡張した形となっている。

2.1.4 複数の項の依存関係を考慮したモデル

最近では、一般化の技術と同時に、周辺文脈をどの程度見て述語と項の一般化を行うか、という点も重要な課題とされている。研究の初期段階では、「述語の意味は動詞のみで決まる」という前提が置かれていたが、これでは述語の主たる意味が項に依存する場合（「実施する」等の機能動詞など）に、選好性を正しく計算できない可能性があった（例えば、「セールを実施する」の主語には店が入る一方で、「国勢調査を実施する」の主語には政府機関が入る）。先に述べたモデルのうち、Ritter ら、Van de Cruys らのモデル [12, 18, 8, 9] は、このような問題に対応するため、主語・動詞、または動詞・目的語の2つ組を述語とみなし、選択選好性知識の一般化を行っている。しかしながら、1節で述べたような、談話内の言及による名詞への付加的な情報が述語の選択選好性に影響を与える問題にはやはり無力である。

2.2 事象間関係知識を用いた談話解析に関する研究

本研究は、先行文脈による項の意味の変化を捉え、その変化を述語の選択選好性の計算に反映させることを目的としている。これは、照応解析の分野で用いられ

ている, Narrative Chain [22] に代表される事象間関係知識 (X attacked → arrested X) [23, 24] と類似している. 事象間関係知識を用いた照応解析では, 先行詞候補と照応詞の参与している事象の間の関連性を計算し, それを手がかりとすることで, 解析の性能が向上することが報告されている [14, 3, 5]. これらの研究で用いられる事象間関係知識において, 項は離散的な意味クラスで表現され, 述語の汎化は行われていない. そのため, 述語の完全一致により事象間の関連性を計算しており, データスパースネス問題がより顕著であると予想される. 一方で, 本研究の枠組みは, 述語と項を分散表現に基づいて表現するため, 未知のテスト事例に対してより頑健であることが期待される. 分散表現を用いて事象間関係知識を獲得する手法として Modi ら [25] の研究があるが, 述語間の関係を捉えることを目的としており, 格関係は考慮していない. つまり, “X attacked Y → police arrested X” と “Y attacked X → police arrested X” の区別をつけていないため, “attack した X が arrest の目的語として自然である” のような知識のモデル化を目的とする本研究とは方向性が異なる.

3 文脈を考慮した選択選好モデル

本研究は談話内での言及による名詞の意味的性質の変化を考慮した選択選好モデルを構築する。構築するモデルは分散表現に基づいた手法となっており、Van de Cruys が提案した、ニューラルネットワークを用いたモデルを拡張したものとなっている。このため、本節では Van de Cruys が提案したニューラルネットワークを用いた手法について説明し、その後提案手法への拡張方法について説明する。

3.1 ニューラルネットワークを用いたモデル

ニューラルネットワークを用いた手法として、Van de Cruys は（動詞，目的語）の組に対する選択選好性のモデル化と、（主語，動詞，目的語）の三つ組に対する選択選好性のモデル化を提案している。ここでは、（主語，動詞，目的語）の三つ組に対する選択選好性モデル（以降 SVO モデル）について紹介する。

SVO モデルは 1 層の隠れ層を持つフィードフォワード型ニューラルネットワークで構成されており、選択選好性の獲得と、項と述語の意味ベクトルの学習を同時に行う。モデルのネットワーク（主語 s ，動詞 v ，目的語 o ）の三つ組に対し、組合せの尤もらしさを表すスコアを計算する。図 1 に (police, arrest, criminal) の三つ組の選択選好スコアの計算例を示す。選択選好性スコア $score(s, v, o)$ は

$$score(s, v, o) = W_2 \mathbf{h} \quad (1)$$

$$\mathbf{h} = \tanh(W_1(V(s) \oplus V(v) \oplus V(o)) + b) \quad (2)$$

と計算する。ここで $V(w) \in \mathbb{R}^d$ は単語 w のベクトル、 $\mathbf{h} \in \mathbb{R}^h$ は隠れ層、 $W_1 \in \mathbb{R}^{h \times 3d}$ 、 $W_2 \in \mathbb{R}^{1 \times h}$ は各層の重み行列、 b はバイアス項である。また、 \oplus はベクトルの連結 (concat) を表している。なお Van de Cruys の提案モデルでは、単語 w が主語，動詞，目的語のいずれかによって異なる単語ベクトルを用いていたが、役割によらず単一のベクトルを用いたほうが高い性能が得られたため、本研究では単一のベクトル $V(w)$ を用いた。

SVO モデルは前述の通り（主語，動詞，目的語）の組に対し、述語の選択選好性を満たすか否かを識別するモデルである。そのため、訓練事例として、選択選

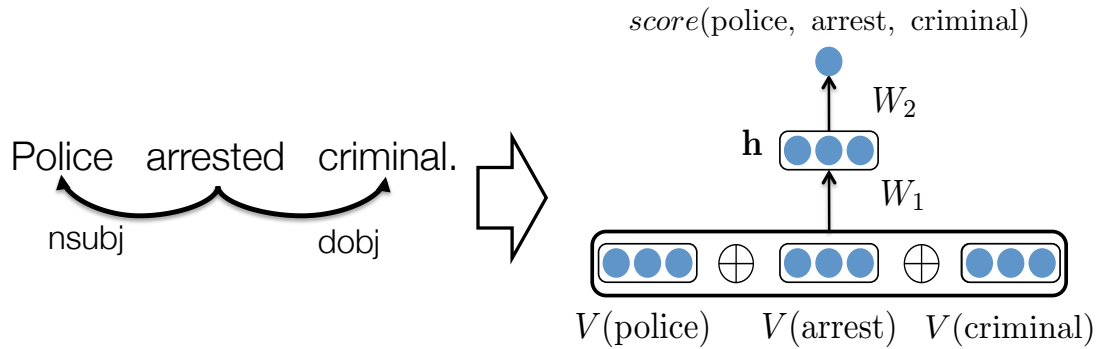


図 1: SVO モデルのネットワーク構造

好性を満たす組（正例）と選択選好性を満たさない組（負例）が必要となる．そこで，コーパス中に出現した述語と項の共起事例を正例として扱い学習を行う．コーパス中に出現する述語と項の共起事例は，人間が記述した自然な文の中での共起であるため，述語と項の組合せは選択選好性を満たしていると言えるといった仮定に基づいている．また，Collobert ら [26] が提案した教師なし学習を用いることで，人手で負例を作成することなく，正例をもとに自動的に擬似負例を作成し学習を行う．具体的な方法としては，まず訓練コーパスから得られた（主語，動詞，目的語）の組を正例とし，この組に対し単語の出現頻度分布に基づき「主語」，「目的語」，「主語と目的語の両方」をそれぞれ別の単語に置き換えた擬似負例を作成する．次に，これらの正例と擬似負例についての選択選好スコアを式 (1) を用いて計算し，式 (3) を最小化するようにモデルパラメータ（単語ベクトル $V(w)$ と重み行列 W_1, W_2 ，バイアス項 b ）を学習する．

$$\begin{aligned}
 & \max(0, 1 - \text{score}(s, v, o) + \text{score}(\tilde{s}, v, o)) \\
 & + \max(0, 1 - \text{score}(s, v, o) + \text{score}(s, v, \tilde{o})) \\
 & + \max(0, 1 - \text{score}(s, v, o) + \text{score}(\tilde{s}, v, \tilde{o})) \quad (3)
 \end{aligned}$$

ここで， \tilde{s} ， \tilde{o} はそれぞれランダムにサンプルし置き換えた主語と目的語である．つまり，コーパスから得られた正例に対するスコアと，ランダム置換により生成した擬似負例に対するスコアの差が 1 以上になるようにモデルの最適化を行う．

表 1: 本研究で扱う先行文脈

| 先行文脈内の述語の種類 t | 例文 |
|-----------------|--|
| 他動詞 (主語) | <i>John attacked Bob</i> , so the police arrested <i>John</i> . |
| 他動詞 (目的語) | Tom report <i>John</i> , so the police arrest <i>John</i> . |
| 自動詞 | <i>John cheat</i> , so police arrest <i>John</i> . |
| 形容詞 | <i>John is suspicious</i> , so police arrest <i>John</i> . |
| コピュラ | <i>John is suspect</i> , so police arrest <i>John</i> . |

3.2 項の先行文脈を考慮した選択選好モデルの構築

1 節で述べたとおり，本研究は談話内での言及による名詞の意味的性質の変化を考慮した選択選好モデルを設計する。

今，例として 1 節の例文 (1) と同様の問題を解いていると仮定する。

(1) [*John_(i)*] attacked [*Bob_(j)*]. Police arrested *him_(i)*

本研究の目的とする文脈による付加要素を考慮した選好性の計算は，例文 (4a), (4b) のように，先行文脈の代わりに名詞が関係節等によって修飾されている文を考え，SVO モデルにおける項ベクトルの部分に，この修飾関係を表現した分散表現ベクトルを埋め込んで選好性を評価することだと考えると直感的に分かりやすい。

(4) a. Police arrested [*John, who attacked Bob*]

b. Police arrested [*Bob, whom John attacked*]

項の意味的性質に変化を与える修飾関係としては述語表現，副詞表現，同格，A of B など様々な要素が考えられるが，本研究では項の周辺文脈のモデル化への第一歩として，事象間関係を応用した照応解析の先行研究 [14, 3, 5] に倣い，選択選好性の判定対象となる名詞が項として出現する述語（動詞，形容詞，コピュラ）の述語項関係を文脈として採用する。本研究で扱う先行文脈の種類について表 1 にまとめた。

述語項関係を表現するベクトルについては、Hashimoto ら [27] や Modi ら [25] が構成要素から計算する手法を提案しており、本研究もそれに倣い、構成性に基づいて文脈付きベクトルを計算する。本研究の手法では、述語の種類によって名詞と述語の意味的關係が異なると考え、述語の種類毎に異なる合成関数を考える。また、他動詞のように述語が複数の項をとる場合、対象名詞が先行文脈の動詞においていずれの格關係になっているかで名詞が帯びる意味が異なるため、これを区別する。上記の例であれば、例文 (4a) の *John* と例文 (4b) の *Bob* は共に (*John*, *attack*, *Bob*) という先行文脈を持つが、*John* は先行文脈において主語の役割であるため “attack した” *John* を表現し、一方 *Bob* は先行文脈において目的語の役割であるため “attack された” *Bob* を表現する必要がある。

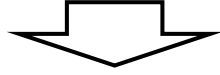
また一般に、文章中には、対象の名詞を項に取り、その名詞の文脈とみなすことのできる複数の述語が書かれていると考えられる。したがって、意味の変化はこれらの重ね合わせとして計算されることが自然だと考えられるが、今回は埋め込みの結果最も選好性の高くなる述語項関係を一つ選択し、この述語項關係が対象名詞の性質に影響を与える文脈であるとする。

このような仮定の結果、本研究が行う文脈の埋め込みは図 2 のようなネットワークで表現される。提案手法では、3.1 節で説明した SVO モデルにおいて、主語ベクトルと目的語ベクトルの部分に、それぞれ単語ベクトルと述語項關係を表現した文脈ベクトルのいずれかを選択できるように拡張する。これにより、対象の名詞に対して先行文脈として述語による言明が存在する場合は文脈ベクトルを用い、文脈がない場合は単語ベクトルを用いることができる。

3.3 定式化

前述のとおり、我々の手法では、述語の種類及び、述語との格關係の違いを区別して文脈付きベクトルの合成を行う。そこで、述語の種類と格關係の組み合わせにより、文脈の種類を { 他動詞-主語, 他動詞-目的語, 自動詞, 形容詞, コピュラ } の 5 種類に分け、文脈タイプに応じた変換行列を用いて単語 w に対する文脈付きベクトルを生成することで、各述語表現における役割を考慮した意味ベクトルの計算を可能にする。

[John] attacked Bob. ... Police arrested [John].



| 主語 | 動詞 | 目的語 |
|--------|--------|--------------------------|
| police | arrest | [John], who attacked Bob |

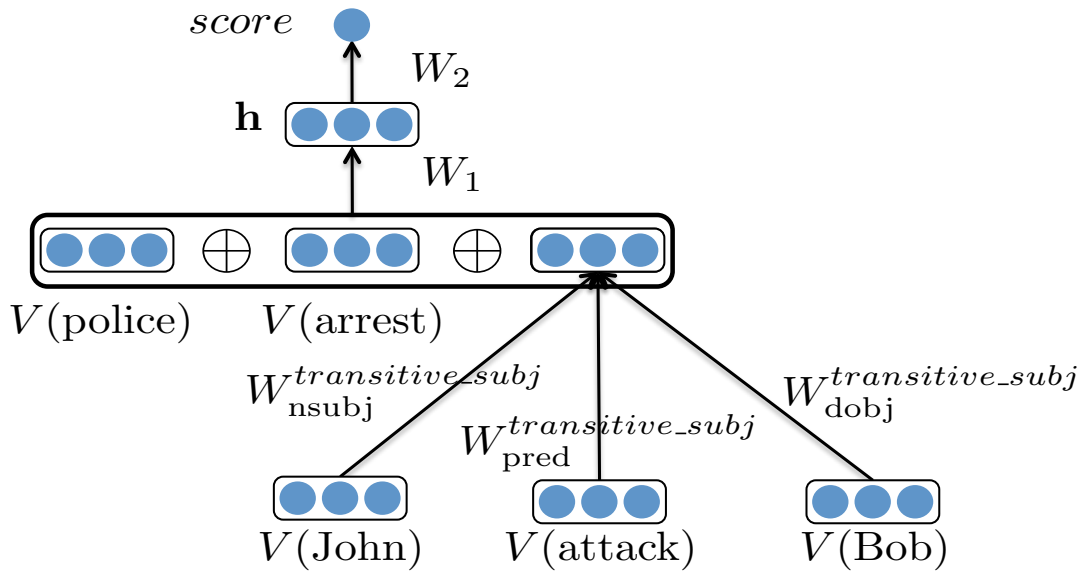


図 2: 提案手法の概要

w に対する先行文脈 c_w による文脈付きベクトルの生成関数 $V'(w, c_w)$ を以下の式で定義する.

$$V'(w, c_w) = \begin{cases} V(w) & \text{if } c_w = \phi; \\ \tanh(W_{\text{pred}}^t V(p) + \sum_i W_{\text{arg},i}^t V(a_i)) & \text{otherwise;} \end{cases} \quad (4)$$

ここで, $W_{\text{pred}}^t \in \mathbb{R}^{d \times d}$ は文脈タイプ t の文脈における述語 p に作用する行列であり, $W_{\text{arg},i}^t \in \mathbb{R}^{d \times d}$ は文脈タイプ t の文脈における項 a_i に作用する行列である. 我々の定義する文脈タイプにおいては, 他動詞の場合には二つの項を取り, それ以外は一つの項を取る. 本研究では, 表 1 に示すように述語の種類と格関係の種類ごとに 5 種類の文脈タイプ ($t = \text{tra_subj}, \text{tra_obj}, \text{intra}, \text{adj}, \text{cop}$) を用いたモデル (Syntactic Category and Gramatical Case Model, 以降 SCGC モデル) と 格関係の

みに着目し 2 種類の文脈タイプ ($t = \text{subj, obj}$) を用いたモデル (Gramatical Case Model, 以降 GC モデル) の二つのモデルを用意した。この定式化により特定の項数を持つ任意の文脈タイプの意味ベクトル計算が可能である。先行文脈が存在しない場合 ($c_w = \phi$) には、単語自身のベクトルを用いる。

提案モデルは、SVO モデルにおける項の入力ベクトル $V(s), V(o)$ を先行文脈を考慮したベクトル $V'(s, c_s), V'(o, c_o)$ に置き換えた (5) で選好性を計算し、従来の選択選好モデルでは捉えることが出来なかった文脈を考慮した選好性の計算を実現する。

$$\max_{c_s, c_o} \text{score}(s, v, o, c_s, c_o) = W_2 \mathbf{h} \quad (5)$$

$$\mathbf{h} = f(W_1(V'(s, c_s) \oplus V(v) \oplus V'(o, c_o)) + b) \quad (6)$$

3.4 訓練事例の作成

本小節では、モデルの学習に用いる訓練データの作成について説明する。

本手法では従来手法がモデル化していた (主語, 動詞, 目的語) の選択選好性と、項に先行文脈による付加情報がついた場合の選択選好性の両方をモデル化する。このため、モデルの学習には

- Type A: (主語, 動詞, 目的語)
- Type B: (主語, 動詞, 目的語, 項の先行文脈となる述語項関係)

の共起事例が必要となる。訓練コーパスに対して以下の処理を行い、これらの訓練事例を獲得する。

1. 動詞・名詞・形容詞の特定及び表記の統一

コーパス中の各単語に品詞タグを付与し、タグをもとに動詞・名詞・形容詞を特定する。また、共起事例を収集する際に名詞の単数形 / 複数形 (policeman / policemen) や動詞の基本形 / 活用形 (arrest / arrested) など単語に表記ゆれが存在すると、それぞれ別の事例として集計してしまい、正しい統計情報を得られない。このため、単語の見出語化 (lemmatize) と小文字への変換を行い、表記を統一する。

2. 述語項関係を特定

単語間の依存構造を表す係り受け関係をもとに述語と、その述語の項を特定する。述語には動詞・名詞・形容詞を用いる。ただし、逆接や否定を表す係り受け関係を持つ動詞・名詞・形容詞は述語として扱わない。逆接や否定を表す係り受け関係を持つ場合“people can not eat table”のように選択選好性を満たさない語を項としてとることが可能となり、モデルの学習の妨げとなる事例が獲得される恐れがあるため除外する。項には述語に対して主語または目的語の関係で係っている名詞を用いる。なお、目的語は動詞の場合のみ項として扱う。この処理により特定された述語項関係のうち、主語と目的語を持つ動詞の述語項関係を Type A の訓練事例として獲得する。

3. 項の先行文脈の特定

獲得された Type A の項（主語または目的語）と共参照関係にある名詞の述語項関係を項の先行文脈として特定し、Type B の訓練事例として獲得する。共参照関係とは、同一のエンティティを指し示す語同士の関係である。本来、共参照関係は句単位で付与されるが、今回は述語項関係との整合性をとるため、句の主辞となる単語間に共参照関係があるとみなす。単語間の共参照関係をもとに以下の二種類の Type B の事例を獲得する。

- Type B1: (主語, 動詞, 目的語, 主語の先行文脈)
- Type B2: (主語, 動詞, 目的語, 目的語の先行文脈)

Type A の主語と共参照関係にある名詞が存在し、その名詞が述語の項となっている場合、Type B1 の訓練事例として獲得する。同様に、Type A の目的語と共参照関係にある名詞が存在し、その名詞が述語の項となっている場合、Type B2 の訓練事例として獲得する。

本研究では、Stanford CoreNLP [28]¹ の解析結果を用いて上述の処理を行う。例文 (5) に対する解析結果を図 3 に示す。

(5) The old man stole money, so the police arrested the man.

¹<http://nlp.stanford.edu/software/corenlp.shtml>

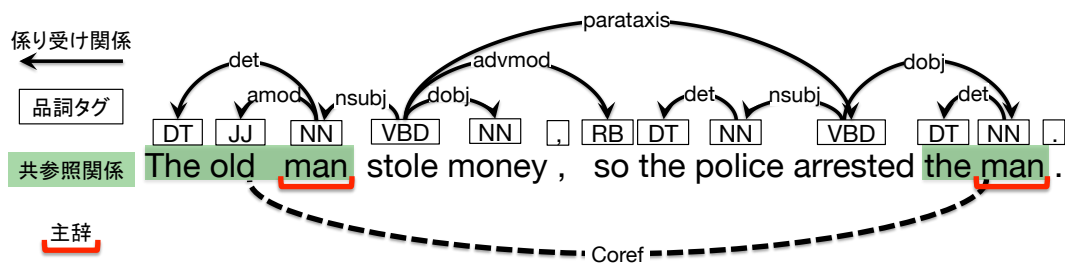


図 3: Stanford CoreNLP による解析結果

Stanford CoreNLP は Penn Treebank タグセット [29] をもとに品詞タグを付与する。今回、動詞・名詞・形容詞として用いた品詞タグを表 2 に示す。

また、係り受け関係は Universal Dependencies [30]²に基づいてラベルが付与される。今回は主語として *nsubj*、目的語として *dobj* を用いる。これにより、図 3 からは、Type A の訓練事例として (man, steal, money) と (police, arrest, man) が獲得される。また、表 3 に示す係り受け関係を逆接や否定を表す係り受け関係として扱う。

上述の通り、共参照関係は図 3 の緑色の部分のように句に対して付与されるが、今回は句の主辞間に共参照関係にあるとみなす。つまり、図 3 では、the old man の man と the man の man の間に共参照関係があるとみなされる。これにより、Type B1 の訓練事例として (man, steal, money, (police, arrest, man, *transitive_obj*)), Type B2 の訓練事例として (police, arrest, man, (man, steal money, *transitive_subj*)) が獲得される。

本研究では、訓練コーパスとして、大規模 Web コーパス ClueWeb12³ の一部 (約 2.2 億文書, 6.9 億文) を用い、Type A, Type B1, Type B2 の訓練事例を獲得し、学習時の正例に用いた。高精度な解析が行える係り受け解析に比べて、共参照解析は対象とする文章によって精度が低くなる。訓練データ作成時の解析ミスによるノイズを可能な限り除くため、同一文内で、かつ表層形が一致する単語間の共参照関係のみを用いた。共参照解析の結果を 100 事例サンプリングし、人手により評価したところ、上記の制約をかけなかった場合の精度が 71% であった

²<http://universaldependencies.org/>

³<http://lemurproject.org/clueweb12/>

表 2: 動詞・名詞・形容詞として用いた品詞タグ

| 品詞タグ | タグの説明 | 本研究での扱い |
|------|-----------------|---------|
| NN | 普通名詞 (単数形) | 名詞 |
| NNS | 普通名詞 (複数形) | |
| NNP | 固有名詞 (単数形) | |
| NNPS | 固有名詞 (複数形) | |
| VB | 動詞 (基本形) | 動詞 |
| VBD | 動詞 (過去形) | |
| VBG | 動詞 (動名詞・現在分詞形) | |
| VBN | 動詞 (過去分詞形) | |
| VBP | 動詞 (現在形：非三人称単数) | |
| VBZ | 動詞 (現在形：三人称単数) | |
| JJ | 形容詞 | 形容詞 |
| JJR | 形容詞 (比較級) | |
| JJS | 形容詞 (最上級) | |

のに対し、解析対象を限定することで精度が 87.0% に向上することを確認した。

訓練コーパスとして用いた ClueWeb12 は Web 文書をクロールして作成されたものであるため、単語の入力ミスやスラングなどが含まれており、必ずしも正しい文で書かれているわけではない。また、英語以外の言語で書かれた文も僅かながら混在している。これらのようなノイズとなる文から獲得された共起事例を取り除くため、コーパス全体で単語の品詞を考慮した出現頻度上位 5 万単語を語彙として用い、これに含まれない単語からなる共起事例は訓練事例から除外した。

訓練事例に出現しなかった単語 (Out of Vocabulary, OOV) に対しても選択選択性の計算を可能にするため、訓練事例中で出現回数が 3 回以下の単語は OOV タグに置換した。

これらの処理を行った結果、Type B1 と Type B2 が合わせて 2,229,624 事例、1,335,505 種類獲得された。また、Type A に関しては上述の処理を行った後、Type

表 3: 逆接や否定を表す係り受け

| 係り受けラベル | 係り先の単語 |
|---------|--------------|
| neg | not |
| neg | never |
| advmod | seldom |
| advmod | rarely |
| advmod | scarcely |
| advmod | hardly |
| advmod | however |
| advmod | nevertheless |
| advmod | nonetheless |
| mark | unless |
| mark | although |
| mark | though |
| conj | but |

B1, Type B2 と訓練事例数を合わせるため, 57 分の 1 をサンプリングし, 2,217,784 事例, 744,029 種類獲得された.

3.5 モデルパラメータの学習

パラメータの学習は, SVO モデルと同様の手順で学習する. ただし, 本手法では以下の拡張を行う.

- 単語ベクトル $V(w)$ と 重み行列 W_1, W_2 , バイアス項 b に加えて, 先行文脈の意味ベクトルの計算に用いる重み行列 $W_{\text{pred}}^t, W_{\text{arg},i}^t$ の学習も行う.
- 学習には, SVO モデルの学習に用いた (主語, 動詞, 目的語) の共起事例と共参照関係から作成した文脈付き事例の両方を用いる.

- 擬似負例として項を置換する際に，訓練事例における出現頻度分布に従って別の項と置き換える．つまり置換対象として，単語だけではなく，文脈付きの項がサンプルされるようにする．
- 先行文脈の意味ベクトルの計算において，活性化関数 \tanh を適用しているため，先行文脈の意味ベクトルの各次元は $(-1,1)$ の範囲に収まる．一方，単語の意味ベクトルには活性化関数を適用していないため，値域の制限はない．本手法では，先行文脈と単語の意味ベクトルの値域を合わせるため，単語ベクトル $V(w)$ にも活性化関数 \tanh を適用する．

単語ベクトル，隠れ層の次元数はそれぞれ $d = 50, h = 50$ に設定し，モデルパラメータは全て標準正規分布から得た値で初期化した．最適化には Adam [31] を用い，1,000 事例ごとのミニバッチ学習を，学習データ全体に対するイテレーション回数を 30 回として行った．

4 評価実験

本研究で構築したモデルにより，項の先行文脈の情報を反映した述語の選択選好性の計算がどの程度正しく行えたかを確認するため評価実験を行った．本節では，まず実験の目的を確認し，様々な実験設定について説明した後，実験の結果について述べる．

4.1 実験設定

本研究の目的は，項自身の意味的性質に加えて，項が談話内で言及されてきた文脈情報を用いて述語の選択選好性を計算するモデルの構築である．

本評価実験では，照応解析の先行研究 [14, 3, 5] から得られた知見より，「代名詞が参与している事象と先行詞が参与している事象には関連がある」ことを利用し，先行文脈を見ることの効果を照応解析性能の向上の度合いで評価する．具体的には，項の先行文脈を用いないで学習した選択選好モデルに比べ，項の先行文脈も考慮した提案モデルを用いることで照応解析性能が向上することを確かめる．

4.1.1 評価タスク

文書中に出現した代名詞に対して先行詞候補が複数提示されている状況で，先行詞候補の集合を先行詞らしさの順に順序付けするランキング問題を考える．評価指標には，一般的にランキングモデルの評価で用いられる Mean Reciprocal Rank (MRR)，Guu ら [32] が用いたランキング問題に対する評価指標 Mean Quantile (MQ)，及び1位に順位付け出来た割合 Accuracy (ACC) を用いる．より具体的に

は、下記の式でスコアを計算する.

$$\text{MRR} = \frac{1}{N} \sum_{i=1}^N \frac{1}{\text{rank}(i)} \quad (7)$$

$$\text{MQ} = \frac{1}{N} \sum_{i=1}^N \frac{|C(i)| - \text{rank}(i)}{|C(i)| - 1} \quad (8)$$

$$\text{ACC} = \frac{\text{rank}(i) = 1 \text{ となった問題数}}{N} \quad (9)$$

$$(10)$$

ここで、 N は問題総数、 $\text{rank}(i)$ は i 番目の問題において式 (5) によって計算されたスコアに基づいてランキングした際の正解先行詞の順位であり、 $C(i)$ は i 番目の問題における先行詞候補集合である.

なお、述語の選択選好性モデルの先行研究では、*pseudo-disambiguation test* と呼ばれるタスクによる評価が一般的である。*pseudo-disambiguation test* では、コーパス中に出現した（主語、動詞、目的語）を正例とし、これのうちの一つの要素をランダムに入れ替えた事例を負例とした、二値分類タスクである。本研究では、より現実的な問題の上で、選択選好性モデルの文脈化の効果を評価するため、代名詞照応解析の問題上での評価を行う。代名詞照応解析の先行研究では、照応解析性能の評価指標として CEAF, BLANC などのさまざまな指標が提案されているが [33, 34, etc.], 本研究の評価は純粹に選択選好性モデルの評価であるため、ランキングレベルの評価に留める。今後の課題として、提案モデルの結果を既存の照応解析器の解析手がかりの一部として用い、既存研究との比較を行う予定である。

4.1.2 評価データの作成

評価コーパスには OntoNotes 5.0 [35] を用いた。OntoNotes は新聞記事、放送原稿、Web ページなど複数ジャンルからなるコーパスで、品詞、構文木、述語項構造、代名詞・名詞句間の共参照関係等が付与されている。OntoNotes に付与されている述語項構造は、係り受け関係に基づいた構造ではなく、PropBank [36] の意味役割に基づいた構造となっている。本実験においては、意味役割に対して表 4 に示すような主語、述語、目的語との対応付けを行った。なお、コピュラ文や形

表 4: PropBank による意味役割と主語, 述語, 目的語との対応

| 意味役割 | V が be 動詞の際の対応 | V が be 動詞以外の際の対応 |
|------|----------------|------------------|
| V | | 述語 |
| ARG0 | | 主語 |
| ARG1 | 主語 | 目的語 |
| ARG2 | 述語 | |

容詞の述語項関係を得るために, 意味役割 V が be 動詞の場合とそれ以外で異なる対応付けを行った. また, 意味役割は句単位で付与されているが, 句の主辞となる語に対して意味役割が付与されているとみなした.

以下では, 下記の談話を用いて, 実験対象となる代名詞および先行詞候補の抽出方法に関して説明する.

(6) In $his_{(\bar{p})}$ 40-minute $speech_{(i)}$, $Chen_{(j)}$ declared the $determination_{(k)}$ of the $people_{(l)}$...(中略)... $Chen_{(j')}$ visited ...(中略), and $he_{(p)}$ stated ... (ectb_1025 改)

評価対象の代名詞として, 提案モデルにより先行詞のランク付けが行える問題, すなわち他動詞の主語もしくは目的語となっている代名詞を対象とした (he_p は対象であるが, $his_{\bar{p}}$ は対象でない). また, 下記の基準により先行詞候補を抽出し, 抽出結果の中に正解の先行詞が存在する問題を評価の対象とした.

- 先行詞候補として, 代名詞の出現した文より前に出現した名詞句の主辞, かつ品詞が**名詞**である (代名詞は含まない) 語を対象とした. 例えば, $he_{(p)}$ に対して, $his_{(\bar{p})}$ **以外**のすべての名詞 i, j, k, l, j' が先行詞候補となる.
- 先行詞候補集合の中に共参照関係にある複数の先行詞候補が存在する場合, 選択選好モデルによる出力値が**最も高い**候補のみを残し順位付けを行った. 例えば, $Chen_{(j)}$ と $Chen_{(j')}$ が共参照関係にある場合, 選択選好モデルの出力値が高い方のみを順位付けの対象とした.
- 訓練事例に出現していない先行詞候補や先行文脈の構成要素は, OOV (Out of Vocabulary) タグに置換して評価した. 例えば, $Chen_{(j)}$ の先行文脈は

$c_j = \langle \text{Chen, declared, determination, subj} \rangle$ であるが、訓練事例の中に declared が現れなかった場合、 $c_j = \langle \text{Chen, OOV, determination, subj} \rangle$ とした。

最終的には、OntoNotes コーパスに出現する代名詞のうち、12,661 問の代名詞を評価対象とした。

4.1.3 比較対象

先行文脈を見ることの効果を見積もるため、ベースラインモデルを構築した。より具体的には、訓練事例のうち (主語, 動詞, 目的語) のみを与えて訓練を行い、先行詞候補のランキング時に項の先行文脈を一切与えないモデルを構築した。このベースラインモデルは、3.1 節で説明した Van de Cruys の SVO モデルに相当する。

また、先行文脈から意味ベクトルを合成する際に、先行文脈を構成する述語の種類と格関係の両方によって合成に用いる重み行列を変化させる SCGC モデルと格関係のみによって変化させる GC モデルの二つのモデルの比較も行う。

4.2 実験結果

4.2.1 先行文脈の考慮による性能変化

提案手法が、先行文脈による項への付加情報を考慮して選択選好性の計算を行っているかを確認するため、選択選好性計算の際に先行文脈を考慮しない従来モデル (SVO モデル) との比較を行った。先行詞候補ランキング問題の MRR, MQ, ACC に基づく評価結果を表 5 に示す。表 5 より、提案モデルは SVO モデルよりも高い性能を達成しており、項の先行文脈の情報を反映した選択選好性の計算が行えていることが確認できる。また、Wilcoxon の符号付き順位検定による有意差検定においても、SVO モデルに対して $p < 0.05$ となり有意差があることも確認した。先行文脈から意味ベクトルを合成する際に、先行文脈を構成する述語の種類と格関係の両方によって合成に用いる重み行列を変化させる SCGC モデルと格関係のみによって変化させる GC モデルとの比較においては、MRR, ACC での

表 5: 先行詞候補ランキング問題での MRR

| 設定 | MRR | MQ | ACC |
|------------------|---------------|---------------|---------------|
| SVO モデル (ベースライン) | 0.2350 | 0.6852 | 0.1228 |
| 提案モデル (GC) | 0.2853 | 0.7560 | 0.1502 |
| 提案モデル (SCGC) | 0.2959 | 0.7558 | 0.1631 |

表 6: 先行文脈を構成する述語ごとの学習事例数

| 先行文脈内の述語の種類 | 学習事例数 |
|-------------|---------|
| 他動詞 (主語) | 639,390 |
| 他動詞 (目的語) | 463,106 |
| 自動詞 | 874,137 |
| 形容詞 | 203,019 |
| コピュラ | 49,972 |

評価においては、SCGC モデルが高い結果となり、MQ では GC モデルが高い結果となった。ただし、この二つのモデルに対して Wilcoxon の符号付き順位検定による有意差検定をおこなったところ有意差は生じなかった。このことから、今回のモデルにおいて先行文脈から意味ベクトルを合成では、先行文脈内の格関係による影響が大きく、述語の種類による影響は小さいと考えられる。述語の種類ごとに異なる重み行列を用いて合成を行う SCGC モデルの方が GC モデルよりも合成による表現力が高く、より高い性能を示すと期待していたが、そのような結果は得られなかった。この原因としては、先行文脈内の述語ごとに学習事例数が異なっていたことが考えられる。先行文脈内の述語ごとの学習事例数を表 6 に示す。表 6 より、形容詞とコピュラの学習事例数が少なく、特にコピュラに関しては自動詞の約 17 分の 1 しかなかった。学習事例数が少なかったためコピュラや形容詞に対して、述語ごとに重み行列を変化させた SCGC モデルで重文に十分に行えていなかったと考えられる。

4.2.2 訓練データ量と性能

次に，学習に用いる文脈情報付き事例の量と，提案モデル（SCGC）の性能との関係について調査した．図 4, 5, 6 は，学習に用いた文脈付き訓練事例の量を 1 100% の間で変化させた際の学習曲線である．なお図 4, 5, 6 では，縦軸がそれぞれの評価指標による性能，横軸が（主語，動詞，目的語）の共起データに加えて使用した文脈付きデータ使用量の割合を示している．図 4, 5, 6 より，使用する文脈付き訓練事例の量を増やすことによるモデルの性能向上を確認でき，データ量増加に伴う性能の飽和が発生していないことも確認できる．従って，現状では学習事例量が不十分であり，事例数の増加により更なる精度向上が期待できる．事例数の増加による精度向上の度合いや飽和点の確認のため使用データ量を増加させて検証する必要がある．

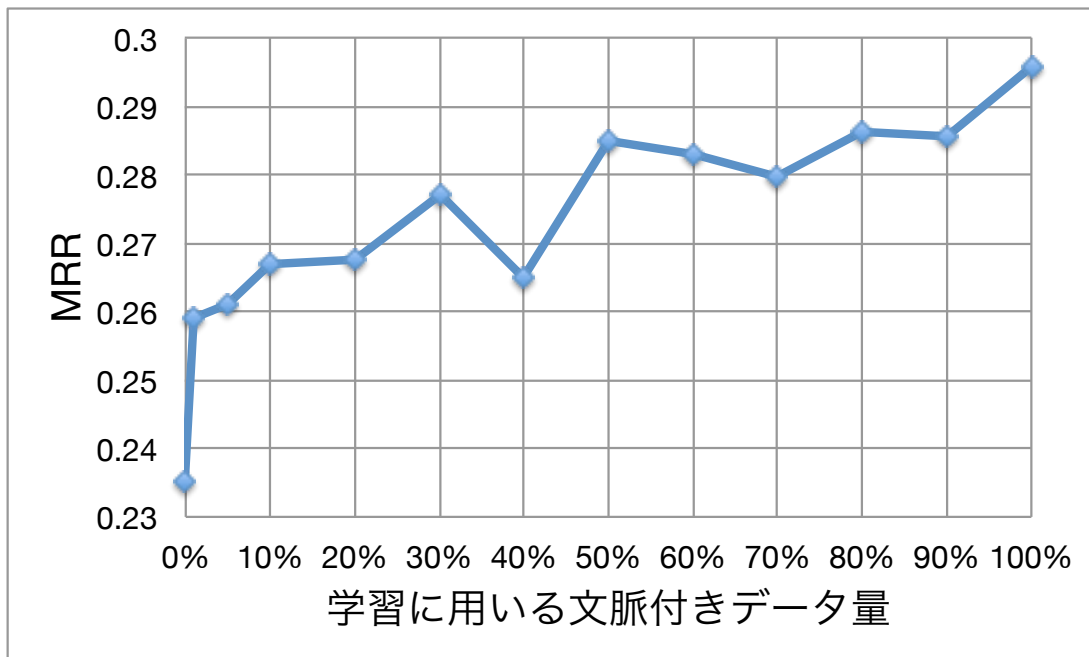


図 4: 訓練に用いる共参照事例数の変化による MRR の変化

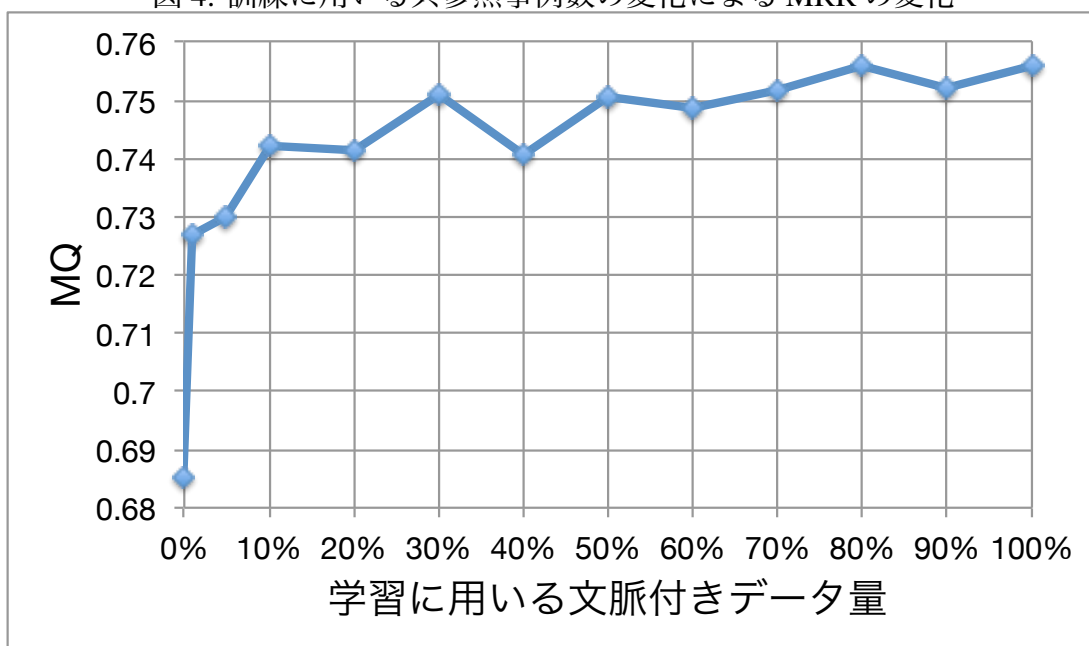


図 5: 訓練に用いる共参照事例数の変化による MQ の変化

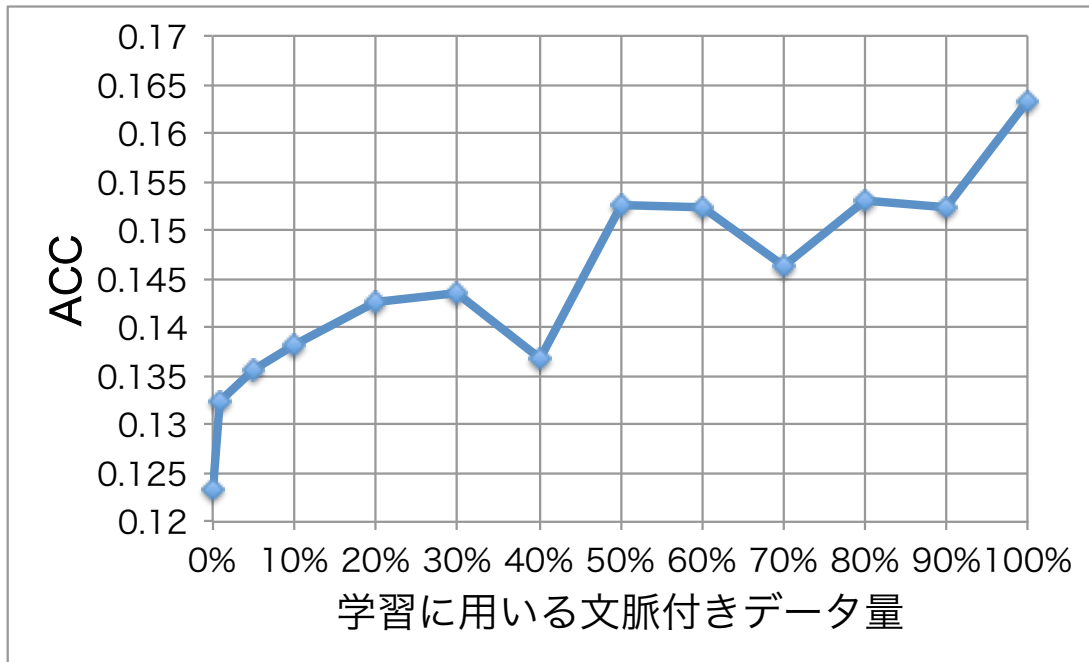


図 6: 訓練に用いる共参照事例数の変化による ACC の変化

4.2.3 単語ベクトルと性能

次に、単語ベクトル及び隠れ層を変化させた場合の性能がどう変化するかを調査した。単語ベクトル d 及び隠れ層 h の次元数を 25, 50, 100 とした場合の性能の変化を表 7 に示す。SVO モデル、提案モデル (SCGC) とともに次元数を変化することで性能も僅かに変化したものの、有意差が生じる程の変化は確認できない程度であり、次元数が性能に大きな影響を与えるものではないことが確認できた。また、単語ベクトルの初期値として外部コーパスで学習した単語の意味ベクトルを用いた場合の性能を比較した結果を表 7 の下部に示す。外部コーパスで学習した単語の意味ベクトルには、Wikipedia と Gigaword Corpus 5 (LDC2011T07) [37] を用いて、GloVe[38]⁴ により学習したベクトルを用いた。表 7 より、SVO モデル、提案モデルともに、単語ベクトルの初期値に外部コーパスで学習した単語ベクトルを用いることで性能向上が確認できる。学習済みベクトルを初期値に用いることで、学習事例中での出現頻度が低い単語に関しても選好性の計算を適切に

⁴<http://nlp.stanford.edu/projects/glove/>

表 7: 単語ベクトル及び隠れ層の次元数を変化させた際の結果

| 設定 | MRR | MQ | Acc |
|-----------------------------|---------------|---------------|---------------|
| SVO モデル d=25, h=25 | 0.2356 | 0.6835 | 0.1217 |
| SVO モデル d=50, h=50 | 0.2350 | 0.6852 | 0.1228 |
| SVO モデル d=100, h=100 | 0.2378 | 0.6803 | 0.1288 |
| 提案モデル d=25, h=25 | 0.2849 | 0.7573 | 0.1527 |
| 提案モデル d=50, h=50 | 0.2959 | 0.7558 | 0.1631 |
| 提案モデル d=100, h=100 | 0.2740 | 0.7397 | 0.1434 |
| SVO モデル d=50(GloVe), h=50 | 0.2420 | 0.6891 | 0.1273 |
| SVO モデル d=100(GloVe), h=100 | 0.2448 | 0.6893 | 0.1324 |
| 提案モデル d=50(GloVe), h=50 | 0.3015 | 0.7539 | 0.1693 |
| 提案モデル d=100(GloVe), h=100 | 0.2889 | 0.7458 | 0.1568 |

行えるようになったと考えられる。

4.3 事例分析

提案モデルの振る舞いを確認するために、提案モデルにより正解先行詞の順位が改善された 100 事例を、人手により分析した。改善が確認された事例の中には、項の先行文脈で言及された事例と、代名詞を項に持つ述語の間に関連性があり、提案モデルによりそれらの関連を適切に捉えることができたと考えられる事例が存在した。以下に、その一例を示す。

(7) (前略) ... *Rodney Sutton*_(i) broke a seven-year-old world record by shearing 839 lambs in nine hours. A crowd of hundreds watched *him*_(pro) accomplish the feat. ... (cnn_0160)

ここで、解析対象となる代名詞は *him*_(pro) であり、正しい先行詞候補は *Rodney Sutton*_(i) である。SVO モデルによるランク付けでは 13 位であったが、提案モデルによるランク付けでは 1 位に改善された。SVO モデルでは、*him*_(pro) の述語である「偉業を成し遂げる (accomplish the feat)」の主語に *Sutton* が入りやすいか、といった基準でランキングが行われており、*Sutton* は人であるため選択選好性がある程度満たすが、人を表す他の候補との差は無かった。一方、提案モデルでは、「記録を破った *Sutton* (*Sutton* broke record)」が「accomplish feat」の主語になりやすいか、という尺度でランキングが行われ、選好文脈による付加情報を考慮したうえで選択選好性を満たしているかを計算することで、他の候補よりも適切であると判別できたと考えられる。

本実験では、表 3 に示す、逆接・否定を表す語と係り受けにある述語の項となる代名詞は評価対象から除外してあるが、述語の意味が反転するような修飾語は他にも多く考えられる。例えば、“Tom answered (incorrectly)” という文において、副詞 *incorrectly* の有無によって、述語の意味が大きく変わってしまう。このため、逆接や否定語を予め除外するのではなく、提案モデルの入力形式を拡張し、述語の主語・目的語以外の項（例えば、in restaurant などの前置詞句）などと共に本論文で捨象している情報を扱える枠組みを構築する必要があると考えられる。

5 おわりに

本論文では、述語の選択選好性モデルを談話解析に用いる場合に特に重要となる、項の先行文脈を考慮した述語の選択選好性計算モデルを提案した。述語の選択選好性モデルに関する先行研究では、項となる名詞自身の意味的性質に基づいて選好性の学習・計算が行われていたのに対し、本研究では、先行文脈における項に対する言及から、構成的に項の意味を組み上げる分散表現ベースの枠組みを提案した。提案モデルを先行詞候補のランキング問題により評価・分析した結果、提案した枠組みにより、正しい先行詞候補を上位にランキングできるようになることを確認した。

今後の課題として、提案モデルの結果を既存の照応解析器に組み込み、解析手がかりの一部として用いて、既存の解析手法との比較を行うことが挙げられる。ただし、提案モデルでは他動詞の項となっている代名詞以外を扱うことが出来ないため、現状のモデルを既存の解析手法に組み込んだとしても、効果は限定的なものだと考えられる。そのため、他動詞以外の述語の選択選好性も計算できるモデルに拡張する必要がある。また、本研究では、項の先行出現文脈として共参照関係を頼りに同一指示語を項にもつ述語項関係を用いたが、それ以外にも副詞・否定詞による述語への修飾など、選択選好性を考える上で重要な周辺文脈要素は多数存在しており、これらに対応することでさらにモデルの性能が向上することが期待される。また、本論文では、ただひとつの述語項構造を項の先行文脈として扱ったが、1節で述べたように、先行文脈における項に対する言及は一般的には複数存在する。このため、時間的順序や依存関係を考慮できる Recursive Neural Network のような分散表現モデルを用いて、先行文脈の複数の言及の内容や、言及の順序を反映した上で項の意味計算を行うような機構を構築も今後の課題である。

謝辞

本研究を進めるにあたり、多くの方々にご協力をいただきました。ここに、心より感謝の意を表します。

主指導教官である乾健太郎教授には、お忙しい中、研究活動全般にわたり温かいご指導、ご助言をいただきました。心より感謝を申し上げます。同じく、研究内容について多くのご助言をいただきました岡崎直観准教授に深く感謝します。ご多忙の中、審査委員をお引受けくださいました、周暁教授、大町真一郎教授に深く感謝致します。本研究を進めるにあたり、数々の的確なご助言をくださり、研究活動を暖かく指導してくださいました、井之上直也助教に心より感謝致します。研究方針や手法に関する数々のご指導ご助言をくださいました松林優一郎研究特任助教に深く感謝致します。最後になりますが、研究生生活の様々な場面でお世話になりました研究室の皆様有難うございました。

参考文献

- [1] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational linguistics*, Vol. 28, No. 3, pp. 245–288, 2002.
- [2] Shane Bergsma. Discriminative Learning of Selectional Preference from Unlabeled Text. In *EMNLP*, pp. 59–68, 2008.
- [3] A. Rahman and V. Ng. Resolving Complex Cases of Definite Pronouns: The Winograd Schema Challenge. In *Proceedings of EMNLP-CoNLL*, pp. 777–789, 2012.
- [4] Benat Zapirain, Eneko Agirre, Lluís Màrquez, and Mihai Surdeanu. Selectional preferences for semantic role classification. *Computational Linguistics*, Vol. 39, No. 3, pp. 631–663, 2013.
- [5] Haoruo Peng, Daniel Khashabi, and Dan Roth. Solving Hard Coreference Problems. In *NAACL*, pp. 809–819, 2015.
- [6] Philip Resnik. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, Vol. 61, No. 1, pp. 127–159, 1996.
- [7] Katrin Erk, Sebastian Padó, and Ulrike Padó. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, Vol. 36, No. 4, pp. 723–763, 2010.
- [8] Tim Van de Cruys, Thierry Poibeau, and Anna Korhonen. A tensor-based factorization model of semantic compositionality. In *HTL-NAACL*, pp. 1142–1151, 2013.
- [9] Tim Van de Cruys. A neural network approach to selectional preference acquisition. In *EMNLP*, pp. 26–35, 2014.
- [10] Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. Inducing a semantically annotated lexicon via em-based clustering. In *Proceedings*

of the 37th Annual Meeting of the Association for Computational Linguistics, pp. 104–111, College Park, Maryland, USA, June 1999. Association for Computational Linguistics.

- [11] Diarmuid O Séaghdha. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 435–444. Association for Computational Linguistics, 2010.
- [12] Alan Ritter, Oren Etzioni, et al. A latent dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 424–434. Association for Computational Linguistics, 2010.
- [13] Daisuke Kawahara, Daniel W Peterson, and Martha Palmer. A Step-wise Usage-based Method for Inducing Polysemy-aware Verb Classes. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1030–1040, 2014.
- [14] N. Inoue, E. Ovchinnikova, K. Inui, and J. Hobbs. Coreference Resolution with ILP-based Weighted Abduction. In *COLING*, pp. 1291–1308, 2012.
- [15] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.
- [16] Hang Li and Naoki Abe. Generalizing case frames using a thesaurus and the mdl principle. *Computational linguistics*, Vol. 24, No. 2, pp. 217–244, 1998.
- [17] Daisuke Kawahara and Sadao Kurohashi. Case frame compilation from the web using high-performance computing. In *NAACL*, pp. 1–7, 2006.
- [18] Tim Van de Cruys. A non-negative tensor factorization model for selectional preference induction. *Journal of Natural Language Engineering*, Vol. 16, No. 4, pp. 417–437, 2010.

- [19] Richard Socher, Christopher D Manning, and Andrew Y Ng. Learning continuous phrase representations and syntactic parsing with recursive neural networks.
- [20] Richard Socher, John Bauer, Christopher D Manning, and Andrew Y Ng. Parsing with compositional vector grammars. In *ACL (1)*, pp. 455–465, 2013.
- [21] J.R. Firth. A synopsis of linguistic theory 1930-1955. *Studies in linguistic analysis*, pp. 1–32, 1957.
- [22] N. Chambers and D. Jurafsky. Unsupervised Learning of Narrative Schemas and their Participants. In *ACL*, pp. 602–610, 2009.
- [23] D. Lin and P. Pantel. DIRT: discovery of inference rules from text. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference*, pp. 323–328, 2001.
- [24] T. Aviv J. Berant and J. Goldberger. Global Learning of Typed Entailment Rules. In *ACL*, pp. 610–619, 2008.
- [25] Ashutosh Modi and Ivan Titov. Inducing neural models of script knowledge. *CoNLL-2014*, p. 49, 2014.
- [26] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, Vol. 12, pp. 2493–2537, 2011.
- [27] Kazuma Hashimoto, Pontus Stenetorp, Makoto Miwa, and Yoshimasa Tsuruoka. Jointly learning word representations and composition functions using predicate-argument structures. In *EMNLP*, pp. 1544–1555, 2014.
- [28] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60, 2014.

- [29] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, Vol. 19, No. 2, pp. 313–330, 1993.
- [30] Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. Universal stanford dependencies: A cross-linguistic typology. In *LREC*, Vol. 14, pp. 4585–4592, 2014.
- [31] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [32] Kelvin Guu, John Miller, and Percy Liang. Traversing knowledge graphs in vector space. *arXiv preprint arXiv:1506.01094*, 2015.
- [33] Xiaoqiang Luo. On Coreference Resolution Performance Metrics. In *HLT/EMNLP*, pp. 25–32, 2005.
- [34] M Recasens and E H Hovy. BLANC: Implementing the Rand Index for Coreference Evaluation. *Journal of Natural Language Engineering*, 2010.
- [35] Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pp. 57–60. Association for Computational Linguistics, 2006.
- [36] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, Vol. 31, No. 1, pp. 71–106, 2005.
- [37] Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. *English Gigaword*, fifth edition edition, 2011.
- [38] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, Vol. 14, pp. 1532–1543, 2014.

発表文献一覧

受賞一覧

- 大野雅之, 井之上直也, 松林優一郎, 岡崎直観, 乾健太郎 「言語処理学会第 22 回年次大会 最優秀賞」 2016 年 3 月
- 大野雅之, 井之上直也, 松林優一郎, 岡崎直観, 乾健太郎 「第 225 回自然言語処理研究会 優秀研究賞」 2016 年 1 月

国内会議・研究会論文

- 大野雅之, 井之上直也, 松林優一郎, 岡崎直観, 乾健太郎 「分散表現による文脈情報を用いた選択選好モデル」 言語処理学会第 22 回年次大会, 2016 年 3 月
- 大野雅之, 井之上直也, 松林優一郎, 岡崎直観, 乾健太郎 「分散表現に基づく選択選好モデルの文脈化」 情報処理学会研究報告 Vol.2016-NL-225, 2016 年 1 月
- 五十嵐祐貴, 大野雅之, 岡崎直観, 乾健太郎 「キーワードの自動拡張に基づくイベント言及ツイートの収集」 言語処理学会第 21 回年次大会, 2015 年 3 月