視線情報を用いた述語項構造解析モデルへの単語分散表現の導入

牧 諒亮　　西川 仁　　徳永 健伸

東京工業大学 大学院情報理工学研究科

{maki.r.aa@m, hitoshi@c, take@c}.titech.ac.jp

1 はじめに

近年，機械学習は様々な分野において衆目を集めており，自然言語処理においても重要な手段の一つとなっている．中でも，人間が情報を付与したデータを用いる教師あり学習は，さまざまな自然言語処理タスクにおいて利用されている．人間が情報を付与（アノテーション）したテキストはアノテーション付きコーパスと呼ばれ，さまざまなものが言語資源として公開されている．

現在の教師あり学習の枠組みにおいては，アノテーションの結果として付与された情報が利用されるに留まっている．しかし，アノテーションの結果付与された情報だけでなく，アノテーション過程におけるアノテーション作業者の振る舞いからも解析に有益な情報が得られる可能性がある．アノテーション作業者の振る舞いを解析に利用する手法の一つとして，我々は視線情報を日本語述語項構造解析に導入する方法を提案した (Maki et al., 2016)．本稿では，我々の提案手法に単語の分散表現を導入し，解析性能の改善を図る．

ここで (Maki et al., 2016) で取りあげた例文を検討する．

(1) [私$_{ガ}$] は昨日，友達とランチを食べに行きました．値段にも味にも 満足し ました．

(2) 私は昨日，[友達$_{ガ}$] とランチを食べに行きました．値段にも味にも 満足し てくれたようです．

これらの文章の組は，各文章の 2 文目の末尾が異なることを除いて同じ要素から構成されている．一方，下線を施した述語「満足する」のガ格の項は，(1) では「私」，(2) では「友達」と異なっている．これらの文章を利用して「満足する」に対する訓練事例を生成する際には，(1) の「私」と (2) の「友達」が正例となりそれ以外の候補は負例となる．ここで (1) の「友達」と (2) の「私」に着目すると，これらの候補は負例のなかでも別の文脈では正例となりうる候補である．その意味でこれらの負例は「よい」負例であり，「ニアミス候補」であるといえる．

(Maki et al., 2016) では，正解となる項を除いて作業者が最もよく見ていた候補をニアミス候補として扱いランキング学習を行った．このニアミス候補は作業者を最も惑わせた候補であるということができ，それゆえ正解の候補と意味的に近い関係にあるのではないかと考えられる．たとえば (1) におけるニアミス候補「友達」は，「私」と同じく人間を指しており，他の候補とくらべて意味的に近いと言える．述語項構造解析モデルの構築に際しては，Taira et al. (2008) の研究のように，意味に関する素性は有効にはたらくと考えられる．しかしながら，日本語語彙大系の意味カテゴリを利用した我々の評価実験においてはこの知見に反する結果が得られた．その理由として素性の曖昧性解消を無視した自動付与，語彙のカバー率の低さが要因の一部として考えられた．カバー率の低さの原因の一つに語の認定単位の問題が挙げられる．本稿で使用する光田ら (2014) のアノテーション実験で収集されたコーパスでは複合語 [1] を一つの語として扱っているため，日本語語彙大系でカバーされない語が多くなってしまう．この問題を解決するため，評価実験と単語分割単位を揃えたうえで語の分散表現を学習し，語の意味に関する素性が述語項構造解析に有効にはたらくかどうか，またニアミス候補を導入したモデルにおいて有効にはたらくかどうかについて検討を加える．

2 関連研究

単語の分散表現は，たとえば機械学習での素性として，あるいはニューラルネットにおける語の埋め込みの初期値として，様々な用途に用いられるようになっている．大規模な学習データを扱いやすく，また語彙数が大きくなっても同じ枠組みで扱いやすいことがその背景として考えられる．近年の研究では，学習によって得られた分散表現は加法構成性を備えていることが指摘されており，$\boldsymbol{v}_{king} - \boldsymbol{v}_{man} + \boldsymbol{v}_{woman} \approx \boldsymbol{v}_{queen}$ の例がよく知られている．

単語の分散表現を生成する手法は，語の共起行列などコーパスから得られる統計情報を用いる手法 (Lund and Burgess, 1996) と局所的な文脈ベクトルを利用し

[1] BCCWJ（後述）における長単位

たニューラルネットによる手法 (Mikolov et al., 2013; Mnih and Kavukcuoglu, 2013) に大別できる．これらの手法にはそれぞれ長所・短所があり，それぞれの短所を補うモデルとして考案されたのが GloVe (Pennington et al., 2014) である．GloVe は語の共起をもとに最小二乗法によって分散表現を学習する手法であり，skip-gram (Mikolov et al., 2013) など他の手法を上回る性能を備えていることが報告されている．

3 単語の分散表現と意味カテゴリ

3.1 語の認定

本研究で扱う文書は，すべて現代日本語書き言葉均衡コーパス (Balanced Corpus of Contemporary Written Japanese; BCCWJ) (Maekawa et al., 2014) から採取されている．BCCWJ は様々な分野からサンプリングされた約 17 万の文書から構成される約 1 億語のサイズを持つコーパスであり，単語の分割に関しては長単位・短単位という 2 つの基準を設けている．大雑把に言えば，複合語を一つの語として扱うのが長単位であり，複合語を構成する語に分割するのが短単位である．本研究においては，BCCWJ における長単位を単語分割基準として扱うこととした．

3.2 意味カテゴリ

単語の意味カテゴリとして日本語語彙大系を利用した．日本語語彙大系は日本語のシソーラスであり，約 30 万語に対して約 3,000 種類の意味カテゴリが付与されている．Taira et al. (2008) は利用するデータ内の名詞に対して，日本語語彙大系の意味カテゴリのうち，上位から数えて第 3 番目または第 4 番目の階層に位置するカテゴリを人手で付与している．日本語語彙大系の意味カテゴリは一つの語に対して複数のカテゴリが付与されることがあるため，Taira et al. (2008) の意味カテゴリ付与は語義曖昧性解消を含んでいると言える．

本稿においては人的なコストを抑えるため，意味カテゴリを機械的に付与することにした．BCCWJ コアデータ内の各語について，正規化された語形により日本語語彙大系の見出し語を探し，一致するものがあればその語に割り当てられた意味カテゴリを付与する．なお語義曖昧性解消は行っておらず，シソーラスに示されている意味カテゴリを (上位概念も含めて) すべて付与する．またある述語に対して項を予測する際，述語の意味カテゴリが項予測に有効であるとは考えられないが，述語の意味カテゴリと項候補の意味カテゴリのペアであれば項予測に有効であると期待できるため，述語についても意味カテゴリを付与する．

3.3 分散表現

本稿では，BCCWJ における長単位を単語の分割単位として採用し，BCCWJ の約 17 万文書を利用して分散表現を学習した．文分割は BCCWJ の「文頭ラベル」に基づき，約 59 万文からなる約 1 億トークンの訓練用テキストを得た．単語ベクトルの次元は 50 次元と 200 次元の 2 種類とし，ウィンドウサイズを 15，ベクトル化する語の最低出現回数を 5 としたうえで GloVe[2] によりそれぞれ学習を行った．結果として，いずれも 38.5 万トークンについての分散表現を得ることができた．ここで未知語は一つのトークンに代表され，一つの分散表現が割り当てられている．

4 評価実験

4.1 タスク設定

本稿では，光田ら (2014) の設定を踏襲し，(Maki et al., 2016) と同一の設定を採用した．項が述語の前方に出現しているものの中から述語[3] を一つ選択し，アノテーション対象の述語とする．次に述語に先行する項候補を列挙し，述語と各項候補を入力とする．表層格それぞれについて項候補の中から一つを出力し，その項候補が実際に項であれば正解とする．

本研究では特にガ格を対象とする．ガ格は省略が起こりやすく，様々な位置関係に現れるためである．また以下では，ガ格項と述語が同一文内に現れる課題を文内ガ格課題，そうでない課題を文間ガ格と呼ぶ．

4.2 停留の検出

本稿では (Maki et al., 2016) と同じく，光田ら (2014) にならって Dispersion-Threshold Identification アルゴリズム (Salvucci and Goldberg, 2000) を利用し，距離閾値 D を 16 ピクセル，時間閾値 T を 100 ミリ秒として停留を抽出した．さらに停留の座標を停留に含まれる注視点の重心座標によって定義し，停留と項候補を対応づけた．これにより，アノテーション作業者が各候補をどの程度見ていたか，停留回数と停留の持続時間によって評価することが可能となる．

4.3 項候補の順位づけ

(Maki et al., 2016) と同じ方法を利用し，正解の項を 1 位，正解以外で停留回数の最も多い候補を 2 位，停留回数が 0 回の候補を同順を許して 3 位とする．各アノテーション作業者について，各アノテーション課題が 1 つのランキングを構成するように停留をもとに項候補を順位づけたうえで，ランキング SVM (Joachims, 2002) の訓練データとして利用し，モデルのパラメータ推定を行う．視線データはモデルのパラメータ推定

[2] http://nlp.stanford.edu/projects/glove/
[3] ここでの「述語」にはサ変名詞等，事態性名詞も含まれている．

にのみ利用されるため，視線データの存在しない新規の文章に対しても解析を行うことが可能となる．

4.4 モデルの構築

視線データ　評価実験のための視線データとして，日本語を対象とした述語項関係アノテーションの際に光田ら (2014) によって収集されたものを用いる．本稿では，セッション単位でエラー率と停留の数を計上し，視線計測状況が良好なセッションでの視線データを利用することにした．ここでセッションとは 1 人の作業者による 1 課題についての計測開始から終了までの時間区間を指し，エラー率とは，各セッションにおいて 60 Hz で記録される注視点の計測回数を分母，注視点計測に失敗した回数を分子として計算されたものを指している．この過程により 3,680 セッションのうち 2,333 セッションを利用することとした．

アノテーション用文章　光田らのデータ収集で使われた文章は，BCCWJ から採取されており，長単位を単語分割基準として項候補を認定している．アノテーション課題として利用された 221 課題のうち，(Maki et al., 2016) と同じく，文内ガ格の 107 課題と文間ガ格の 77 課題，あわせて 184 課題を訓練データとして利用することとした．

また評価データとして，(Maki et al., 2016) で BCCWJ から作成された 29,519 課題を利用した．このうち 21,816 課題が文内ガ格の課題であり，7,703 課題が文間ガ格の課題である．これらの文章はいずれも BCCWJ のコアデータから取得したものであるため，BCCWJ-DepParaPAS (Maekawa et al., 2014; 小西ら , 2013; 浅原, 大村, 2016; 浅原, 松本, 2013; 植田ら, 2015) から述語項構造アノテーションを利用することができる．

解析器の構築　本稿では，評価のために 2 種類の学習ベースのモデルを用意した．

- **2 値ランキング (Binary)**:
  各訓練事例は人手でアノテーションされたガ格のみを 1 位とし，残りはすべて 2 位とする．
- **停留順ランキング (Fixation)**:
  各項候補に対する停留回数を用い，4.3 に記した方法で順位づけを行う．訓練事例はセッション別に作成されるため，1 課題について作業人数と同数のランキングが生成される．

各モデルの実装には SVM$^{\text{rank}}$ (Joachims, 2006) を利用し，L1 正則化項を加えたうえで学習を行った．いずれのモデルも各項候補のガ格らしさを推定しているため，推定値が最も大きい候補をガ格として出力する．その出力がコーパスのアノテーションと一致した場合のみ「正解」と判定し，正解の項と共参照の関係にある候補であってもコーパスのアノテーションと異なっていれば不正解として扱うこととした．

表 1: 意味カテゴリ素性・分散表現のカバー率

データ	素性	述語	項候補	ガ格項
訓練	φ_{sem}	.83	.36	.60
	φ_{GloVe}	.99	.90	.89
評価	φ_{sem}	.73	.31	.57
	φ_{GloVe}	.92	.84	.82

4.5 素性設計

本稿では分散表現・意味カテゴリ素性の効果をみるため，素性のバリエーションを 3 種類用意する．すべてに共通する素性を ϕ_{base} とし，それに意味カテゴリを加えた素性を ϕ_{sem}，分散表現を加えた素性を ϕ_{GloVe} とする．語の分散表現を除いては意味カテゴリ素性を含め，(Maki et al., 2016) で利用したものと同じものを利用している．

また，分散表現素性については GloVe によって学習された n 次元（$n = 50,\ 200$) の分散表現を用い，入力された語に対応するベクトルが存在しない場合には零ベクトルあるいは未知語ベクトル '<unk>' のいずれかを割り当てることとした．

意味カテゴリ・分散表現素性を除いた素性ベクトルが $\varphi(p,\ c)$ によって生成されるとすると，素性のバリエーションは，それぞれ

$$
\begin{aligned}
\phi_{base}(p,\ c) &= (\varphi(p,\ c)) \\
\phi_{sem}(p,\ c) &= (\varphi(p,\ c),\ \varphi_{sem}(c),\ \varphi_{sem}(p,\ c)) \\
\phi_{GloVe}(p,\ c) &= (\varphi(p,\ c),\ \varphi_{GloVe}(p),\ \varphi_{GloVe}(c))
\end{aligned}
$$

により生成される．ここで $p,\ c$ はそれぞれ述語，項候補を指す．また，意味カテゴリ素性と分散表現のカバー率は表 1 に示したとおりである．ここで「ガ格項」の列は，項候補のうち正解（= ガ格項）が未知語にならない割合を表している．

4.6 結果と考察

各モデル・素性セットの評価結果を表 2 に示す．モデルと素性セットの組合せの精度を比較すると文内ガ格では Fixation モデルの ϕ_{base} が最も高く，文間ガ格では Binary モデルの ϕ_{sem} が最も高い．文間ガ格の精度は全体として低く，うまく解析できてない．文間タスクの方が難しいこともあるが，学習したモデルが文内の候補を優先する傾向にあるため，文間の精度が低くなっていると考えられる．

表 2: 評価実験の結果

モデル	素性	未知語	Accuracy 文内ガ格	文間ガ格
Binary	ϕ_{base}	–	.597	.016
	ϕ_{sem}	**0**	.512	**.055**
	$\phi_{GloVe(50)}$	**0**	.596	.016
		<unk>	.587	.016
	$\phi_{GloVe(200)}$	**0**	.597	.016
		<unk>	.592	.016
Fixation	ϕ_{base}	-	**.621**	.015
	ϕ_{sem}	**0**	.582	.020
	$\phi_{GloVe(50)}$	**0**	.612	.016
		<unk>	.608	.016
	$\phi_{GloVe(200)}$	**0**	.613	.017
		<unk>	.605	.017

次に意味カテゴリ，分散表現の効果について検討する．まず意味カテゴリ素性については ϕ_{base} と ϕ_{sem} を比較すると意味カテゴリを導入することによって文間ガ格の精度が向上しているが，文内ガ格では大幅に精度が低下している．未知語の出現については，零ベクトルを用いた場合と未知語ベクトルを用いた場合を比較すると零ベクトルを用いる方が高い精度が得られている．しかしながら，零ベクトルを用いても分散表現がベースラインの精度を改善しているわけではない．

そこで意味カテゴリと分散表現の素性の影響を調べるため，各モデルについて述語とガ格の係り受け関係ごとに解析精度を調べた．Binary モデル，Fixation モデルのいずれのモデルにおいても意味カテゴリあるいは分散表現の意味的な素性を加えることで，直接係り受け関係にある文内ガ格については精度が下がり係り受け関係にない文内ガ格については精度が向上するという傾向があった．この傾向から，意味的な素性によって，係り受け関係がない場合には意味を手がかりに精度が向上する場合があるが，全体としては精度が低下してしまっているといえる．

また，意味カテゴリ素性の文間ガ格の精度に与える影響が，Binary モデルと Fixation モデルでは大きく異なるなど，現時点で説明のできない現象が残っている．今後はこれらの分析をさらに進めていきたいと考えている．

5 結論

本稿では視線情報を用いた述語項構造解析に対して，意味に相当する素性として語の分散表現の導入を試みた．しかしながら，その結果は解析の精度向上につながるものとはならなかった．この結果に対してはさらなる分析が必要であると思われる．

参考文献

Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2002)*, pages 133–142, 2002.

Thorsten Joachims. Training Linear SVMs in Linear Time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2006.

Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208, 1996.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written Japanese. *Language resources and evaluation*, 48(2):345–371, 2014.

Ryosuke Maki, Hitoshi Nishikawa, and Takenobu Tokunaga. Parameter estimation of Japanese predicate argument structure analysis model using eye gaze information. In *Proceedings of the 26th International Conference on Computational Linguistics*, pages 2861–2869, 2016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

Andriy Mnih and Koray Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in Neural Information Processing Systems*, pages 2265–2273. 2013.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

Dario D. Salvucci and Joseph H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications (ETRA 2000)*, pages 71–78, 2000.

Hirotoshi Taira, Sanae Fujita, and Masaaki Nagata. A Japanese predicate argument structure analysis using decision lists. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 523–532, 2008.

小西 光, 小山田 由紀, 浅原 正幸, 柏野 和佳子, 前川 喜久雄. BCCWJ 係り受け関係アノテーション付与のための文境界再認定. 第 3 回コーパス日本語学ワークショップ予稿集, pages 135–142. 国立国語研究所, 2013.

光田 航, 飯田 龍, 徳永 健伸. 単一述語項関係アノテーション課題における視線情報の収集と分析. 情報処理学会第 217 回自然言語処理研究会, pages 1–8, 2014.

浅原 正幸, 大村 舞. BCCWJ-DepParaPAS:『現代日本語書き言葉均衡コーパス』係り受け・並列構造と述語項構造・共参照アノテーションの重ね合わせと可視化. 言語処理学会第 22 回年次大会発表論文集, pages 489–492, 2016.

浅原 正幸, 松本 裕治. 『現代日本語書き言葉均衡コーパス』に対する係り受け・並列構造アノテーション. 言語処理学会第 19 回年次大会発表論文集, pages 66–69, 2013.

植田 禎子, 飯田 龍, 浅原 正幸, 松本 裕治, 徳永 健伸. 『現代日本語書き言葉均衡コーパス』に対する述語項構造・アノテーション. 第 8 回コーパス日本語学ワークショップ予稿集, pages 205–214. 国立国語研究所, 2015.