

文脈を考慮した前置詞誤り訂正に向けた 前置詞とその潜在意味関係の同時解析

三田 雅人
奈良先端科学技術大学院大学
mita.masato.mz2@is.naist.jp

水本 智也
東北大学
tomoya-m@ecei.tohoku.ac.jp

松本 裕治
奈良先端科学技術大学院大学
matsu@is.naist.jp

1 はじめに

日本人英語学習者が共通して間違いやすい文法項目として、日本語には存在しない「前置詞」が挙げられる。前置詞の用法は複雑であり、意味的に類似した文脈での使い分けに困難を伴うことが多い。日本人大学生によって書かれた英作文からなる学習者コーパス、Konan-JIEM コーパスにおいて前置詞誤りは3番目に多い誤りであり、文法全体の約14%を占めている。このような背景から、前置詞誤りを検出・訂正する研究が盛んに行われている [2, 5, 1]。

前置詞誤り訂正は、多クラス分類問題としての定式化が主流である。これは、個々の訂正候補の前置詞をひとつのクラスとみなし、訂正箇所の前置詞周辺の文脈を用いて適切な前置詞のクラスへと分類するものである。しかし、前置詞がもつ多義性によって、表層上では同じ前置詞でも文脈に依存して多様な潜在的な意味関係を表現することがある。例えば、英文 “I stay in New York” と “We met in 1985” における下線部の “in” は表層上では同じでも、それぞれ「場所」、「時」のような異なった潜在的な意味関係を表現している。

しかしながら、前置詞誤り訂正の既存手法では、前置詞の潜在意味関係を明示的に考慮していない。表層上は同じだが潜在意味関係が異なる場合、各意味関係ごとに出てくる文脈が異なるが、分類器は一つのクラスとして一緒に学習してしまう。

本稿では、前置詞の語義曖昧性解消の Shared Task [3] で用いられた前置詞の語義付きデータを用いることで、前置詞の多義性を考慮した学習手法を提案し、前置詞誤り訂正における前置詞の語義曖昧性の影響について調査および分析を行なう。また、前置詞の語義曖昧性を考慮することで、言語教育への応用に向けた一考察を行なう。

2 前置詞の語義曖昧性解消に関する研究

前置詞の語義曖昧性に関する研究は The Preposition Project (以下, TPP) [4] やそれに関連する前置詞の語義曖昧性解消の共通タスクである SemEval 2007 Shared Task を通じて行われてきた。TPP では 334 個の前置詞に対し、Oxford Dictionary of English に記述されている語義に基づき 847 個の前置詞語義が定義された。SemEval 2007 Shared Task では、TPP で定義された語義をもとに、34 個の前置詞を対象に語義が付与したデータセットを使用した。

Srikumar and Roth [6] は、SemEval 2007 Shared Task で付与された語義を意味的に類似している 32 個の Semantic Relation ラベルにマッピングした。また、ラベル付きデータを用いた Semantic Relation の自動推定手法を提案した。以下の例を使って、Semantic Relation について説明する。

- (a) live at uncle's house
- (b) go to bed at nine o'clock
- (c) hold at knifepoint

(a) において前置詞 “at” は、係り関係にある単語 live, house によって Semantic Relation は “Location” となり、(b) の “at” の Semantic Relation は “Temporal” となり、(c) の “at” の Semantic Relation は “Instrument” となる。このように同じ前置詞でも異なる Semantic Relation となる。また、以下の例、

- (d) reported on September 26

の前置詞 “on” では、Semantic Relation は “Temporal” となり異なる前置詞でも、同じ Semantic Relation をとる。Semantic Relation と前置詞の関係性の例を表 1 に示す。

表 1: Semantic Relation と前置詞の関係の例

Relation	前置詞
Acitivity	at, in, on, to
Cause	at, for, from, of, with
PartWhole	in, of
Location	about, at, by, from, in, of, on, to
Numeric	at, by, for, of, on, to
Via	by, on
Topic	about, on

3 前置詞と Semantic Relation の同時解析

従来の前置詞誤り訂正は、前置詞の潜在的意味関係を考慮していないため、表層が同じ前置詞を一つのクラスとして学習している。しかし、表層上は同じだが潜在意味関係が異なる場合、その前置詞の周辺単語の出現傾向や意味クラスも異なる。係り関係にある単語の表層やその意味クラスがそれぞれ異なる傾向になるため、各素性に対する重みは意味関係ごとに異なるはずだが、従来手法では同一に学習してしまい適切な重みの推定ができていないと考える。そこで、文脈ごとにより適切な重み推定を行うために、Srikumar and Roth [6] で提案された手法を拡張した前置詞と Semantic Relation の同時解析手法を提案する。

Srikumar らの Semantic Relation 解析手法は多クラス分類問題として解くアプローチである。具体的には、解析対象である前置詞とその係り受け関係にある単語を適切な Semantic Relation のクラスへと分類するものである。Srikumar らは前置詞語義を Semantic Relation にマッピングを行なった SemEval 2007 のデータセットを用いて教師あり学習を行った。

本提案手法では、はじめにラベル付きデータセットを用いて、前置詞と Semantic Relation の結合ラベルを作る。例えば、(a) の文では、“Location” という Semantic Relation を “at_Location” とすることで結合ラベルができる。この操作により、各 Semantic Relation ごとに切り分けて異なるクラスとみなすことができ、前置詞の語義曖昧性が解消されたうえで学習を行なうことが可能となる。

4 評価実験

本節では、前置詞の語義曖昧性を考慮したときの前置詞の解析性能の影響を調査するため、以下の実験を行なった。

表 2: 結合ラベルの推定精度

Model	Accuracy
Srikumar 素性	40.84
Srikumar+前置詞素性	47.00

4.1 実験設定

実験のためのデータセットとして、Semantic Relation ラベルが付与された SemEval 2007 のデータセットを用いた。前置詞は種類は多いが、学習者が使用する前置詞は限られている。本稿では、使用頻度の高い上位 11 個の前置詞を対象を絞って実験を行なう。11 個の前置詞は、“of”, “in”, “for”, “to”, “by”, “with”, “at”, “on”, “from”, “as”, “about” である。これにより、実験に使ったデータの事例数は 17,196 文になる。

前置詞の推定を行う前の予備実験として、提案した同時解析モデルの素性を (1) Srikumar ら [6] で使用された素性 (Srikumar 素性), (2) Srikumar 素性に一般的に前置詞訂正に使用される n-gram や前後の単語などを加えたもの (Srikumar+前置詞素性), の 2 種類を用意して結合ラベルの予測を行なった。表 2 に予測結果を示す。Srikumar+前置詞素性の方が推定精度が高かったため、以降の本実験では Srikumar+前置詞素性を使用する。また、素性は同じだが学習時に表層前置詞を一つのクラスとして扱うモデルをベースラインとした。

4.2 評価方法

評価方法は、10 分割交差検定を行ない各前置詞の推定精度を適合率、再現率、F 値を用いて行なった。また、システム全体の精度として、マクロ平均を用いた。なお、各前置詞における適合率、再現率、F 値およびシステム全体のマクロ平均は以下のようにして求めた。

- 各前置詞に対して

$$\begin{aligned} \text{適合率} &= \frac{\text{正しく対象前置詞の推定ができた単語数}}{\text{対象前置詞がついた出力結果の単語数}} \\ \text{再現率} &= \frac{\text{正しく対象前置詞の推定ができた単語数}}{\text{対象前置詞がついた正解データの単語数}} \\ F \text{ 値} &= \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \end{aligned}$$

- システム全体に対して

$$\begin{aligned} \text{マクロ適合率} &= \frac{\text{各前置詞の適合率の合計}}{\text{前置詞の種類数}} \\ \text{マクロ再現率} &= \frac{\text{各前置詞の再現率の合計}}{\text{前置詞の種類数}} \end{aligned}$$

$$\text{マクロ } F \text{ 値} = \frac{\text{各前置詞の } F \text{ 値の合計}}{\text{前置詞の種類数}}$$

4.3 実験結果とエラー分析

表 3: 各前置詞ごとの結果 (適合率・再現率・F 値). 太字はあるシステムが他のシステムの結果より 1 ポイント以上高いものを表す.

前置詞	ベースライン			同時解析手法		
	適合率	再現率	F 値	適合率	再現率	F 値
of	66.40	80.64	72.83	63.63	81.52	71.48
in	41.84	39.01	40.38	38.24	40.78	39.47
from	42.09	42.72	42.41	43.44	42.20	42.81
to	47.81	46.31	47.05	51.03	44.75	47.68
with	55.53	49.23	52.19	57.68	52.30	54.86
for	51.85	47.80	49.74	58.64	39.72	47.36
on	42.81	38.21	40.38	50.25	31.59	38.77
about	44.18	40.43	42.22	35.95	54.38	43.29
at	45.41	38.93	41.92	46.68	36.67	41.13
by	37.09	34.99	36.01	48.72	28.26	35.78
as	52.20	37.11	43.38	44.59	40.23	42.30
トータル	47.93	45.04	46.23	48.99	44.76	45.90

表 3 に実験結果を示す. 同時解析手法はベースラインに比べてシステム全体の精度で見ると, 適合率は改善されたが, 再現率および F 値はほぼ同程度の精度になった. 同じ実験設定で Semantic Relation を推定したときの実験結果を表 4 に示す. こちらも前置詞推定のとおり傾向が実験結果から確認できる. しかし, 各前置詞の精度を見てみると, 同時解析することにより適合率, または再現率が改善する前置詞もある一方で, 同時解析しないほうがよい前置詞もある.

Semantic Relation の推定精度と前置詞の推定精度結果の関係に着目して分析を行う. 表 6 は同時解析を行ったときの各 Semantic Relation の推定精度である. まず, 全ての Semantic Relation 推定の中で最も再現率の高い “PartWhole” に着目する. この “PartWhole” という Semantic Relation が取りうる前置詞は “in” と “of” のみである [表 1 参照]. ここで, 表 3 を見ると, 前置詞 “in” と “of” はどちらも同時解析をした方がベースラインと比較して再現率が向上していることがわかる. 同様に, 全ての Semantic Relation 推定の中で最も適合率の高い “Via” に着目する. この “Via” という Semantic Relation が取りうる前置詞は “by” と “on” であるが [表 1 参照], その両方とも同時解析の方がベースラインに比べて適合率が高い結果となっていることが確認できる. これらの Semantic Relation における推定精度と前置詞における推定精度の関係は, 同時解析することで精度が改善される前置

表 4: Semantic Relation の推定精度比較

Model	マクロ適合率	マクロ再現率	マクロ F 値
ベースライン	55.81	37.59	41.12
同時解析手法	56.84	37.74	41.18

表 5: 学習者文に対する前置詞推定精度比較

Model	マクロ適合率	マクロ再現率	マクロ F 値
ベースライン	22.81	24.21	19.85
同時解析手法	24.65	22.19	18.76

詞もあれば, 反対に同時解析しないほうがよい前置詞もある事象の一つの原因であると考えられる.

次に, 実応用に向けて, 実際に学習者が書いた文に対して同様の実験を行った. 実験設定は基本的には前述の実験設定と同じである. モデルの学習は SemEval 2007 のデータセット全体で行ない, 評価データとして CoNLL2014 Sharaed Task の評価セットを用いた. 表 5 は CoNLL2014 評価セットにおける前置詞推定精度である. ネイティブの文である SemEval データセットを用いて行なった実験結果よりも全体的に大幅に精度が低下した. ドメインが違うことを考慮しても低い値となっていると考える. 学習者の文は, 前置詞だけでなく前置詞の周辺単語にも誤りが含まれるため, ネイティブの文と学習者の文とはモデルが観測する状態分布が異なることが精度低下の原因の一つである. そのため, 提案する同時解析手法を実際の学習者の文に適用する際は, 学習者の誤り傾向をうまくモデルに反映する必要がある.

5 おわりに

本稿では, 先行研究の前置詞訂正手法では考慮していない前置詞の語義曖昧性に着目し, 前置詞誤り訂正における前置詞の語義曖昧性の影響について調査および分析を行なった. 前置詞の多義性を考慮した学習を行なうために, 前置詞とその潜在的な意味関係の同時解析手法を提案した. ネイティブによって書かれたデータセットを用いた評価実験では, 全体の精度としては, ベースラインと同時解析手法とは適合率で若干の精度向上が確認できたもの, 全体として大きな精度の差は確認できなかった. しかし, 各 Semantic Relation における推定精度と, 各前置詞における推定精度の関係性を分析すると, 高精度で Semantic Relation の推定ができたときには, その Semantic Relation が取りうる前置詞の精度は改善され, 一方で Semantic Relation の予測精度が低いときは, 同時解析するよりも Semantic Relation を考慮しないベースラインの方が精度が上回るという傾向にあることが確認できた.

さらに, 提案手法の実応用に向けて, 学習者の書いた文に対しても同様の実験を行った. 実験の結果, ネ

表 6: 同時解析を行ったときの各 Semantic Relation の推定精度

relation	適合率	再現率	F 値
ObjectOfVerb	54.62	64.72	59.25
Location	41.99	46.61	44.18
PartWhole	66.56	90.62	76.75
Topic	42.83	60.00	49.98
Source	35.48	43.21	38.96
Destination	45.31	38.02	41.34
Cause	62.96	46.33	53.38
Instrument	49.07	47.89	48.48
Possessor	56.76	66.75	61.35
Attribute	54.35	38.55	45.11
Species	53.22	50.76	51.96
Recipient	48.93	54.54	51.59
Agent	49.85	32.94	39.67
PhysicalSupport	38.62	40.70	39.63
Participant/Accompanier	63.53	40.52	49.48
Purpose	44.31	30.87	36.39
Separation	53.47	35.75	42.85
Beneficiary	44.26	17.64	25.23
Temporal	66.66	48.85	56.38
EndState	46.15	30.76	36.92
Numeric	77.89	48.05	59.43
Experiencer	71.42	11.36	19.60
Opponent/Contrast	63.15	55.81	59.25
StartState	56.25	43.26	48.91
Activity	83.33	15.78	26.54
MediumOfCommunication	54.54	14.63	23.07
Other	71.42	6.66	12.19
Manner	66.66	16.90	26.96
Via	100.00	11.76	21.05
ProfessionalAspect	50.00	3.17	5.97

イティブの文で行った実験結果の推定精度よりも、学習者の文に対する推定精度は大幅に低下した。本稿で提案した同時解析手法は、ネイティブの文のみで学習しているため、訂正精度を改善するためには学習者の誤り傾向などをモデルに反映させる必要がある。

最後に言語教育への応用という観点から提案手法についての一考察を行う。従来の前置詞訂正を含めた文法誤り訂正システムは、訂正候補の単語のみを出力しているが、本稿で提案した前置詞の潜在的な意味関係を考慮したモデルは“in_Location”のような前置詞とその Semantic Relation の両方を出力することが可能である。これらは実応用を見据えたとき、2つの利点があると考えられる。1つ目は、前置詞だけでなく Semantic Relation を学習者に提示することで、学習者はシステムの出力結果が「正しい結果」なのか、「誤った結果」なのかの判断の手がかりに利用できる。例えば、“jailed *at/for 12 years” と学習者が for を at に誤っ

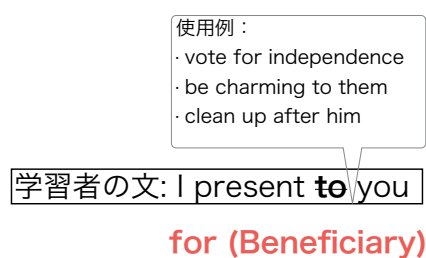


図 1: フィードバックの例

てしまったケースを考える。ここで、システムが“in”と出力した場合、学習者によってはシステムの出力が正しいのか間違いなのか判別がつきにくい場合がある。しかし、in (PartWhole) と Semantic Relation 付きで出力した場合だと、学習者はシステムの出力の間違いに気づきやすくなる。

2つ目は、Semantic Relation と外部知識をルールベースで紐付けることによって学習者にフィードバックが行えることである (図 1)。例えば、各 Semantic Relation にその使用例を事前に紐付けておけば、システムはただ単に誤った前置詞を訂正するだけでなく、学習者に対してその文脈で使用される別の表現を同時に提示することが可能となる。

参考文献

- [1] M. Chodorow, M. Gamon, and J. Tetreault, “The Utility of Article and Preposition Error Correction Systems for English Language Learners: Feedback and Assessment,” *Language Testing*, vol.27, no.3, pp.419–436, 2010.
- [2] R. De Felice and S.G. Pulman, “A Classifier-Based Approach to Preposition and Determiner Error Correction in L2 English,” *Proceedings of COLING*, pp.169–176, 2008.
- [3] K. Litkowski and O. Hargraves, “SemEval-2007 Task6: Word-Sence Disambiguation of Prepositions,” *Proceedings of the Fourth International Workshop on Semantic Evaluation (SemEval-2007)*, pp.24–29, 2007.
- [4] K. Litkowski and O. Hargraves, “The Preposition Project,” *Proceedings of ACL-SIGSEM Workshop*, pp.171–179, 2005.
- [5] A. Rozovskaya and D. Roth, “Algorithm Selection and Model Adaptation for ESL Correction Tasks,” *Proceedings of ACL-HLT*, pp.924–933, 2011.
- [6] V. Srikumar and D. Roth, “Modeling Semantic Relations Expressed by Prepositions,” *Transactions of the Association for Computational Linguistics*, vol.1, pp.231–242, 2013.