Vol.2010-DBS-151 No.39
2010/11/13

口語文書の解析精度向上のための助詞落ち推定および補完手法の提案

池田和史† 柳原正† 服部元† 松本一則† 小野智弘†

評判解析や文書の要約、検索などを高精度に行うために、係り受け解析や格解析が用いられるが、ブログや電子掲示板上の文書を対象とする場合、口語的な記述が多数見られるため、十分な解析精度が得られないことが課題となる。本稿では、口語的な記述に頻繁に見られる助詞落ち表現が解析精度低下の原因の 1 つであることに着目し、助詞落ちを自動的に推定し、欠落した助詞を補完することで解析精度を向上する手法を提案する。提案手法では、新聞などの助詞落ちの少ない正規の文書から意図的に助詞落ちを発生させた文書を正例、助詞落ちを発生させていない文書を負例として識別器を学習させ、解析対象である口語文書の助詞落ち箇所を推定する。加えて、推定した助詞落ち箇所の前後の単語をキーとして新聞文書を検索することで、適切な助詞を自動的に補完する。性能評価実験では、Web から収集したブログ文書に対して、人手により助詞落ち箇所と補完すべき助詞を付与し、提案手法における助詞落ち推定精度および補完精度の評価を行った。加えて、助詞を補完することによる係り受け解析精度の向上についても評価した。

Estimation and Complementation Approach for the Analysis of the Omission of Postposition on Colloquial Style Sentences

Kazushi Ikeda† Tadashi Yanagihara† Gen Hattori†
Kazunori Matsumoto† Chihiro Ono†

In this paper, we propose algorithms for reducing the errors of the dependency analysis on colloquial style sentences by complementing the omission of postpositions which makes dependency analysis errors. In our algorithms, the omission of postpositions is detected by a classifier which is trained by the features extracted from formally written documents such as newspaper sentences. As positive examples of the classifier, we automatically omit the postpositions from newspaper sentences, and as negative examples, we used the newspaper sentences as they are. After estimating the omission of the postpositions, complementation candidates of the omitted postpositions are automatically retrieved from newspapers. In the experimental evaluations, we collect blog documents which contain colloquial style sentences and manually labeled the omitted postpositions on them. We evaluated the estimation accuracy, complementation accuracy, and improvement of the dependency analysis accuracy.

1. はじめに

近年、インターネットの普及により、一般ユーザによる Web 上での情報発信の手段としてブログや電子掲示板が注目されており、これらを対象とした評判解析や要約、検索などに関する研究が盛んに行われている。評判解析などを高精度に行うためには単純な単語の出現頻度や共起に基づく方法だけでなく、単語同士の係り受け関係や主格、目的格といった格情報を利用することが有効とされるが、ブログ文書は口語的な記述を多く含むため、解析精度が低下することが課題となる。係り受け解析や格解析の精度を低下させる原因の 1 つとして、「ラーメン食べた」(「ラーメン(を)食べた」)のような助詞落ち表現が挙げられる。「ラーメン食べた」では、「ラーメン」は「食べた」の目的格として係ることが正しいが、係り受け解析器にかけると「ラーメン食べた」という 1 つの文節として誤った解析が行われる場合などがある。口語的な表現を多く含む文書として、Twitter 上の文書を対象とした著者らの分析では、分析対象の 10,000 文に助詞落ち箇所は 5,087 箇所見られ、係り受け解析器 CaboCha による解析では、そのうち 2,735 箇所(53.8%)に係り受け解析誤りが見られた。

本稿では、これらの助詞落ち表現に着目し、解析対象となる口語文書における助詞落ち箇所を自動的に推定し、欠落した助詞を補完することで解析精度を向上する手法を提案する。提案手法では、新聞などの助詞落ちの少ない正規の文書から意図的に助詞落ちを発生させた文書を正例、助詞落ちを発生させていない文書を負例として識別器を学習させ、解析対象である口語文書の助詞落ち箇所を推定する。加えて、推定した助詞落ち箇所の前後の単語をキーとして新聞文書を検索することで、適切な助詞を自動的に補完する。

提案手法を実装し、性能評価実験を実施した。性能評価実験では Twitter から収集した 10,000 文に対して人手により助詞落ち箇所と補完すべき助詞を付与し、提案手法における助詞落ち推定精度および補完精度の評価を行った。加えて、提案手法によって助詞を補完する前後における係り受け解析の精度向上についても評価した。

†（株）KDDI 研究所, KDDI R&D Laboratories Inc.

2. 関連研究

ブログや電子掲示板を対象とした評判解析や要約、検索などの研究は盛んに行われている 1), 2), 3)。文献 1)はインターネット上の電子掲示板に投稿された口コミ情報などを対象とした評判解析システムのフレームワークを提案している。文献 2)では、電子掲示板の投稿文書に含まれる単語と掲示板のトピックとの関連性に基づき、トピックと関連性の低い発言を取り除くことで要約を自動生成する手法を提案している。文献 3)においては、ユーザが過去に投稿した文書に含まれるキーワードに基づいて、ブログをランキングすることで検索を支援するシステムを提案している。

評判解析や要約の分野において、係り受け解析や格解析を用いることで評判解析や要約の性能が向上するといった知見は文献 4)や 5)などで報告されている。係り受け解析や格解析精度の向上のための研究は古くから行われており、CaboCha 6)や KNP 7)といった優れた言語解析器が提供されているが、新聞文書などの文語的な記述を対象として開発されてきた背景から、ブログや電子掲示板などの口語表現や未知語を多く含む文書では解析精度が低下することが課題である。

係り受け解析や格解析の精度を向上させる方法の 1 つとして、形態素解析精度の向上が有効である。ブログや電子掲示板に対する形態素解析精度向上のための研究も盛んに行われている。口語的表現や話し言葉を言語的な観点などから分析した形態素解析精度向上のための手法として、文献 8) では、「～しちゃう」などの口語特有の言い回しを分析し、人手により辞書登録を行うことで、口語の形態素解析精度が向上することが報告されている。また、形態素解析における未知語の解消については、Web から新語を獲得する手法 9)や未知語の品詞推定を行う手法 10)などが提案されている。著者らはブログ上のくだけた表現を対象とした自動修正手法を提案し、形態素解析精度が向上することを確認した 11)。

係り受け解析精度の向上に関する研究として、文献 12)では、最大エントロピー法に基づくモデルを利用することで、係り受け解析精度を向上させる手法を提案しているが、係り受け解析結果が付与された学習用の文書を要するため、人手によるラベル作業が必要な点が課題である。著者らの予備実験においては、ブログや電子掲示板の文書中には助詞落ち表現が頻繁に見られ、これらが解析精度を低下させる要因の 1 つであることを確認した。助詞落ちに関する研究報告として文献 13)では、助詞落ちや倒置表現の傾向について分析し、解析精度を向上するためのヒューリスティックなルールを提案しているが、人手により生成したルールは作業者が参考にした文例に依存してしたり、主観に基づきやすく、汎用的なルールの生成は困難である。また、文献 14)は口語文書を文単位に分割し、さらに節と呼ばれる細かい単位に分類することで、係り受け解析精度を向上する手法であるが、本稿で対象とする助詞落ちによる係り受け解析誤りは短文においても頻繁に見られるため、文献 14)の手法は最適ではない。

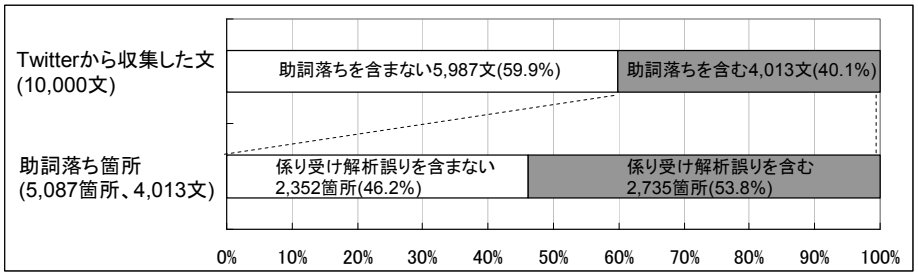

図 1　解析対象文における助詞落ちと係り受け解析誤りの割合

3. 助詞落ちの分析

本稿で対象とする助詞落ち表現について分析する。助詞落ち表現を多く含む文書として、本稿では Twitter から特定のテーマに関して述べられた文書をハッシュタグを指定して収集した。特定のテーマに対する投稿を対象として評判解析や要約、検索を高精度に行うことはマーケティングなどに応用可能な重要な技術といえる。なお、提案手法は Twitter 上の文書に限らず、ブログや電子掲示板など一般ユーザが投稿する文書全般に応用可能な汎用的な手法である。

Twitter から収集した文書に対し、URL や返信を表す”RT”などの記号、ハッシュタグなどを除いた後、「。」、「！」、「？」などの記号を区切り文字として文単位に分割し、8 文字以上の日本語を含む文章を 10,000 文を用意した。提案手法では、助詞落ちが発生している箇所を推定し、文法的に正しい助詞を補完することで、係り受け解析などの精度を向上する。そのため、Twitter から収集した 10,000 文に対して、人手によって次の 3 つのラベルを付与した。(1)助詞落ちが発生している箇所、(2)助詞落ち箇所が係り受け解析誤りを含むかどうか、(3)助詞落ち箇所に補完可能な文法的に正しい助詞。

(1)の助詞落ち箇所と(2)の助詞落ち箇所における係り受け解析誤りについて、人手によるラベル付けの結果を図 1 に示す。本稿で対象とする助詞落ち箇所は、文献 13)や 15)などを参考に、主要な助詞である「が、を、に、で、の、は、と」の 7 種類のいずれかが欠落している箇所とした。分析の結果、助詞落ち箇所は 10,000 文中に 5,087 箇所存在した。文単位では 4,013 文が 1 つ以上の助詞落ちを含んでおり、Twitter 上の文書には多数の助詞落ちが含まれることが確認された。次に、助詞落ちが発生していた 5,087 箇所のうち、係り受け解析に誤りが発生していた箇所は 2,735 箇所(53.8%)であり、助詞落ち箇所において係り受け解析誤りが発生しやすい傾向が確認された。

表 1　助詞落ち箇所と補完可能な助詞の具体例

本文(“/”は助詞落ち箇所を表す)	補完可能な助詞
この荷物 / 至極軽そうだな	が、は
でもそろそろバイト / いくにゃうぅ。	に
仕方なくどら焼 / 購入なう	を
コレ / 何どういう企画ｗｗｗ	は
伊勢海老のお味噌汁 / 美味しそうです。	が、は
誰 / 目線ですか（笑）	の
渋滞だからといってエアコン / 止めるのは、難しいよ。	を
あのカンジ / いい～！	が、は
カード / 浮いてたけど？	が、で、は、と
昆虫グミも最近 / 話題だけど、こっちは本物だし	の、は

表 2　助詞ごとの補完可能な助詞落ち箇所数と割合

補完可能な助詞	が	を	に	で	の	は	と	計
箇所	1,032	1,130	49	98	102	2,138	2,233	6,782
比率(%)	15.2	16.7	0.7	1.4	1.5	31.5	33.0	100

(3)の助詞落ち箇所に文法的に正しい助詞を補完する場合、投稿文書のみから投稿者の意図を完全に把握することは困難であり、正解を一意に定義することが難しい場合がある。例えば「ラーメン食べたかった。」という投稿文書があったとき、「ラーメン」と「食べたかった」の間に補完すべき助詞は「が」や「を」が適切と思われるが、文書に現れていない投稿者の状況まで考慮すれば「は」や「と」が正しい場合もある。これらのあいまい性は評判解析、要約、検索など係り受け解析や格解析を利用するアプリケーションレベルで吸収できるものと考え、本稿では各文において文法的に不自然でない助詞を正解としてラベル付けを行った。すなわち上記の「ラーメン食べたかった」に対しては「が」、「を」、「は」、「と」は正解とし、「に」、「で」、「の」は不正解とした。ラベル付与結果の具体例を表 1 に示す。また、解析対象の 10,000 文において、補完対象とする 7 種類の助詞ごとに集計した結果を表 2 に示す。「は」や「と」が多く、次に、「が」、「を」が多いことが分かった。助詞落ち箇所に補完可能な助詞は 1 箇所あたり平均 1.3 個であった。

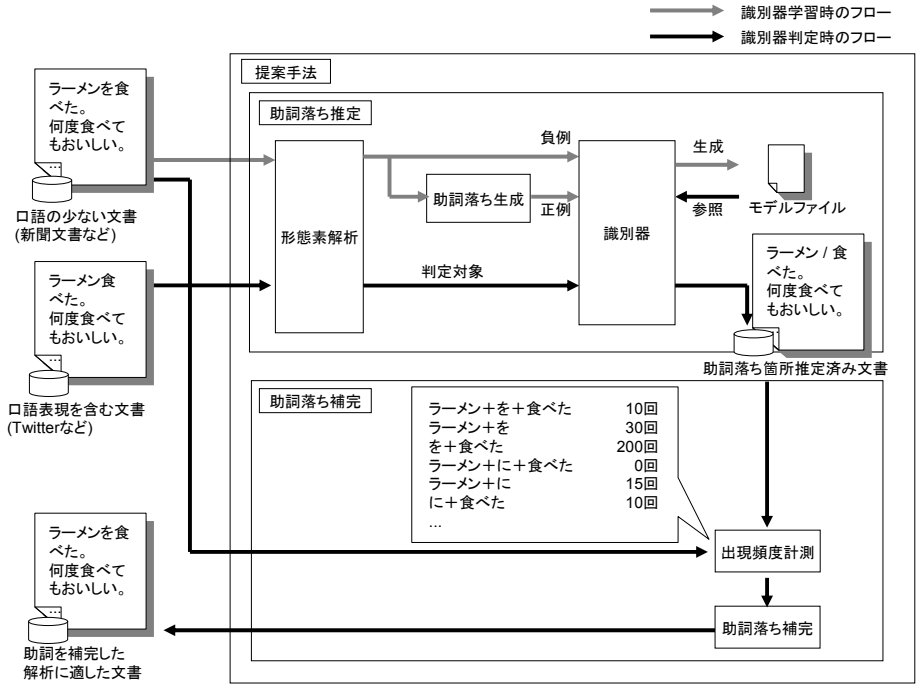

図 2　提案手法の概要

4. 提案手法

提案手法の全体像を図 2 に示す。提案手法は助詞落ち箇所を推定する手法と推定した助詞落ち箇所に対して助詞を補完する手法からなる。

4.1 助詞落ち箇所の推定

助詞落ち箇所の推定方法について述べる。著者らの調査では助詞落ち箇所を推定し、適切な助詞を補完するような手法はこれまでに見つかっておらず、助詞落ち箇所を推定するためのルールは未知といえる。人手によりルールを記述することはできるが、作業者が参考にした文例に依存したり、主観に基づきやすい。例えば、「ラーメン食べた」という助詞落ちを含む文例などから、「名詞の直後に動詞が現れる場合は助詞落ちである」というルールを定義しても「何度食べてもおいしい」のような文例を助詞落ちとして誤検出してしまうなど、高精度に検出することは難しい。

新聞文書から取り出した文:
ラーメンを食べていた。

形態素解析結果:
ラーメン 名詞,一般,*,*,*,*,ラーメン,ラーメン,ラーメン
を 助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
食べ 動詞,自立,*,*,一段,連用形,食べる,タベ,タベ
て 助詞,接続助詞,*,*,*,*,て,テ,テ
い 動詞,非自立,*,*,一段,連用形,いる,イ,イ
た 助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
。 記号,句点,*,*,*,*,。,。,。 ・・・(1)

助詞「を」削除し、「を」の前後4品詞列を正例の特徴量として抽出:
文頭
名詞,一般,*,*,*,*
動詞,自立,*,*
助詞,接続助詞,*,*,*,*
動詞,非自立,*,*
助動詞,*,*,* ・・・(2)

動詞「食べ」と助詞「て」の間には助詞落ちがないため、その前後4品詞列を負例の特徴量として抽出:
文頭
名詞,一般,*,*,*,*
助詞,格助詞,一般,*,*,*
動詞,自立,*,*,
助詞,接続助詞,*,*,*,*
動詞,非自立,*,*,
助動詞,*,*,*,
記号,句点,*,*,*,* ・・・(3)

図 3 助詞落ち推定用学習データの生成手順

提案手法では、助詞落ちを含む文書を正例、助詞落ちを含まない文書を負例として識別器を学習させて用いることで、助詞落ち箇所の高精度な推定を試みる。識別器の学習には一般に、正しくラベル付けされた大規模なデータを必要とするが、助詞落ち箇所がラベルされた大規模な文書を用意することは困難であることから、新聞文書から自動的に助詞落ちを発生させた文書を正例とすることで識別器を学習させる。

4.1.1 新聞文書を用いた助詞落ちの学習

識別器に与える学習データの生成方法について図 3 に具体例を示して説明する。学習用の新聞文書から取り出した各文に対して形態素解析を行い、品詞の並びを得る(図 3 の(1))。ここでは形態素解析器として MeCab 16)を用いた。品詞が助詞であるものを取り除いた品詞列を正例、取り除いていない品詞列を負例として識別器を学習させる。正例は助詞落ち箇所の前後 N 個の品詞を、負例は任意の品詞間を中心とする前後 N 個の品詞を特徴量として与える。本稿では、実験による観測から N=4 とした。ただし、文頭や文末に達した場合は、それぞれを表す記号を与えるものとした。具体的には、図 3 の(2)では、助詞「を」を取り除き、「を」が現れる前後 4 品詞列を抽出して正例としている。図 3 の(3)では、「食べ」と「て」の間の助詞落ちではない箇所を中心とした前後 4 品詞列を抽出することで負例を生成している。特徴量として品詞の並びを与えることで、単語の並びを与えるよりも識別器の汎化性能を高めることができる。提案手法では正例を自動的に生成するため、この例では助詞「て」を取り除いた品詞列も同様に正例として生成されるが、「ラーメンを食べいた。」という非文の品詞列が口語文書に出現する頻度は低いと考えられ、助詞落ち表現を優先的に検出できることが期待される。また、図 3 の例のように正例と負例は同じ文書から生成しても構わないが、本稿では正例と負例を異なる新聞文書から生成した。

4.1.2 口語文書における助詞落ち箇所の推定

解析対象となる口語文書の各文に対して、学習データの生成と同様に形態素解析を行い、任意の品詞間に助詞落ちがあると仮定して、前後 4 品詞列を特徴量として識別器に与えることで助詞落ちの有無を推定する。ここで、助詞落ち箇所の推定精度を向上するために、まず、再現率が高くなるようなパラメータを利用して助詞落ち箇所を過剰に検出しておき、4.2 節で説明する助詞落ちの補完手法において、適切な補完候補を見つけられない場合に助詞落ち箇所ではないと判定することにより助詞落ち箇所を絞り込むことで、助詞落ち箇所の推定と補完精度をそれぞれ向上する。

4.1.3 助詞落ち箇所の推定精度評価の予備実験

3 節において分析に利用した Twitter から収集した 10,000 文に対して提案手法により助詞落ち箇所を推定し、その再現率と適合率を求めた。識別器として SVM(Support Vector Machine) 17)を用いた。内部パラメータである Cost 値および Gamma 値については学習に利用した新聞文書を用いて最適化した。SVM 内部で計算される判定の信頼度を表す Probability 値に対し、助詞落ちがあると推定する閾値を変化させたときの再現率(Recall)、適合率(Precision)の関係を図 4 に示す。再現率、適合率は下記のように定義した。比較のための単純なルールとして、名詞の直後であり、かつ次の品詞が助詞でない箇所を全て助詞落ちと推定した場合の精度は再現率 88.0%、適合率 19.3%であったことから、提案手法による助詞落ち箇所の推定は単純なルールと比べて有効性があるといえる。

再現率 = 助詞落ちを正しく推定した箇所数 / 評価対象文中の総助詞落ち箇所数 **(1)**

適合率 = 助詞落ちを正しく推定した箇所数 / 助詞落ちと推定した総箇所数

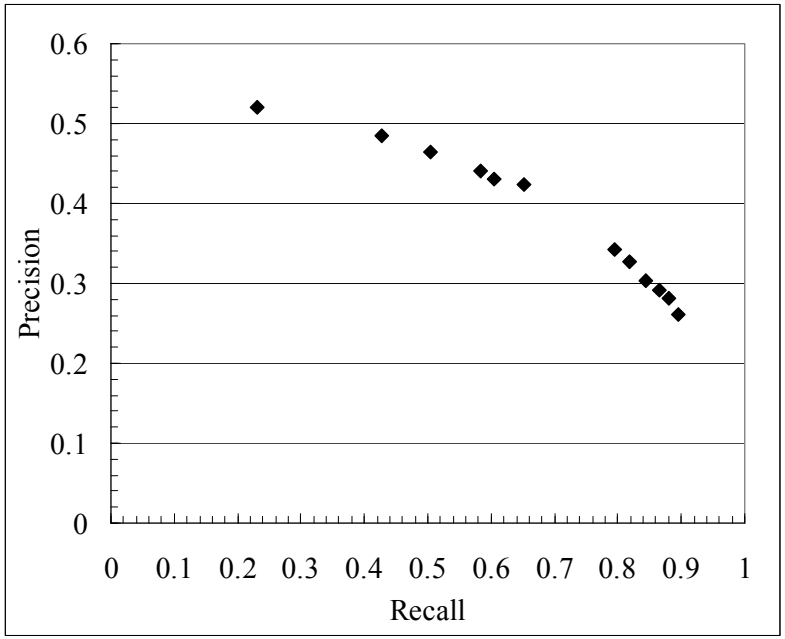

図 4　助詞落ち箇所の推定精度

4.2 助詞落ち箇所の補完

助詞落ち箇所の補完においては、(1)文法的に最も正しいと思われる助詞 1 つを補完する、(2)文法的に正しいと思われる助詞全てを補完の対象とする、という 2 つの補完方法が考えられる。正しい係り受け関係を求める上では(1)で十分であるが、評判解析や要約、検索などで文法的に正しい複数の格情報を取得したい場合は(2)が必要となる。提案手法は補完候補となる助詞を文章が自然となるような尤度の高い順にスコアリングすることで、(1)、(2)の両方に適用可能である。

提案手法では、助詞落ち箇所に補完すべき助詞の候補を新聞文書から検索する。補完手順の詳細について図 5 に具体例を示して説明する。「ラーメン食べた」という文において、「ラーメン」と「食べた」の間に助詞落ちが発生していると推定したとき、「ラーメン」と「食べた」の間に入るような助詞を新聞文書から検索するためのクエリを生成する。クエリは助詞落ち箇所を 1 文字または 0 文字のワイルドカード(”?”で表す)とし、助詞落ち箇所に隣接する前後最大 4 形態素と合わせて検索し、補完対象の助詞「が、を、に、で、の、は、と」および助詞が入らない場合、のそれぞれと一致する

助詞落ち推定済み文:
　ラーメン / 食べた。

新聞文書に対するクエリの生成:
　ラーメン？食べた。
　ラーメン？食べた
　ラーメン？食べ
　ラーメン?
　?食べた。
　?食べた
　?食べ　　・・・(1)

新聞文書の検索結果:

ラーメン食べた。	0回
ラーメンが食べた。	0回
ラーメンを食べた。	0回
...	
ラーメン食べた	0回
ラーメンが食べた	3回
ラーメンを食べた	10回
...	・・・(2)

クエリ長と検索結果の件数から補完候補となる助詞をスコアリング:

候補文	スコア
ラーメンを食べた。	832
ラーメンは食べた。	428
ラーメンが食べた。	219
...	
ラーメン食べた。	64
...	・・・(3)

図 5　助詞落ちの補完手順

件数を求める。図 5 の(1)では助詞落ち箇所の前の形態素は最大 1 つで「ラーメン」であり、後の形態素は最大 3 つで「食べ」、「た」、「。」となる。これらを組み合わせて 7 個のクエリを生成している。新聞文書における各クエリの検索件数は図 5 の(2)のようになる。検索結果はより長いクエリに一致するほど、文脈を反映していると考えられ、正しい補完が行われると期待される。そのため、クエリに用いた形態素数と検索件数を用いて候補となる助詞をスコアリングする(図 5 の(3))。助詞落ちを補完しない場合よりも補完した場合の方がスコアが高く、かつ閾値以上のスコアが得られた場合にのみ助詞の補完を行うことで、過剰適用を抑制する。

5. 性能評価実験

5.1 実験環境と手順

提案手法の性能を評価するための実験の手順と環境を示す。助詞落ち補完精度の評価では、提案手法を用いて助詞落ち箇所を推定、補完したときの(1)助詞落ち箇所の推定精度と(2)助詞の補完精度、(3)助詞落ち補完前後の文の係り受け解析の精度向上について評価した。(1)は予備実験として実施した助詞落ち推定手法に加えて、新聞文書の検索結果を用いて過剰適用を抑制したときの助詞落ち箇所の推定精度について評価した。(2)は助詞落ち箇所の推定が正しく、かつ補完した助詞も正しい場合を正解としたときの精度について評価した。(3)については、助詞落ちを補完した文をサンプリングして係り受け解析の精度が向上したかどうかを人手により評価した。

以下に実験環境の詳細を示す。形態素解析器は MeCab を用い、係り受け解析には CaboCha を用いた。

- 形態素解析器：MeCab Version 0.97
- 形態素解析辞書：MeCab 標準 IPADIC 辞書にブログ等で頻出の未知語 30 万語を追加
- 係り受け解析器：CaboCha Version 0.53
- 係り受け解析モデルファイル : CaboCha 標準の京大コーパスから学習したモデルファイルを利用
- 提案手法で利用する新聞文書 : 毎日新聞（2003 年～2009 年） 600 万文。助詞落ち推定の学習データとして正例 50 万文、負例 50 万文。補完候補となる助詞を検索するために 500 万文を利用
- 評価対象の口語文書 : 特定のテーマに対する Twitter 上の投稿から収集した 10,000 文

5.2 助詞落ち推定精度の評価

4.1 節の予備実験では、識別器の判定信頼度に対して助詞落ちと推定する閾値の取り方によって再現率、適合率が異なることを確認した。提案手法では、4.1 節の SVM による助詞落ち箇所の推定において、再現率が高くなるようなパラメータを用いて助詞落ち箇所を過剰に検出し、4.2 節の助詞補完のスコアに対する閾値を調整することで過剰適用を抑えられることから、これら 2 つの閾値を変化させたときの再現率、適合率について評価を行った。実験結果を図 6 に示す。F 値が最大となるのは再現率 66.7%、適合率 74.9%のときであった。また、再現率 24.5%のとき、適合率 89.2%の高精度で助詞落ち箇所を推定することができた。

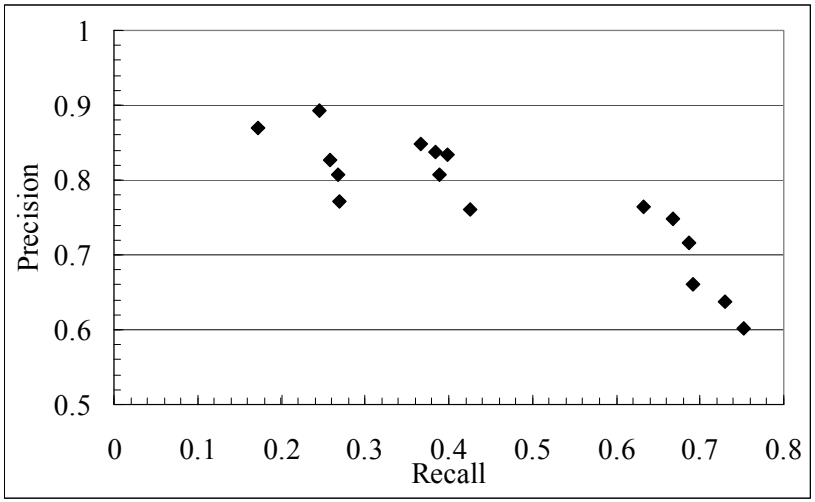

図 6 助詞落ち箇所の推定精度

5.3 助詞落ち補完精度の評価

次に、助詞補完の正しさについて評価する。ここでは 2 通りの方法で精度を評価した。(1)文法的に最も正しいと思われる助詞 1 つを補完する。(2)文法的に正しいと思われる助詞全てを補完対象とする。ただし、各助詞落ち箇所における補完可能な助詞数は与えられるものとした。例えば、「が」と「は」を補完することが正解であるような助詞落ち箇所においては、提案手法で補完する助詞数も最大で上位 2 件までとした。再現率、適合率はそれぞれ次のように定義する。(1)は助詞落ち箇所数に基づくのに対し、(2)は補完した件数に基づいている。

(1)の場合の定義 **(2)**

再現率 = 助詞落ちを正しく補完した箇所数 /総助詞落ち箇所数

適合率 = 助詞落ちを正しく補完した箇所数 /助詞落ちと推定した総箇所数

(2)の場合の定義

再現率 =助詞落ちを正しく補完した件数 /助詞落ち箇所に補完可能な総助詞数

適合率 =助詞落ちを正しく補完した件数 /助詞落ちを補完した総件数

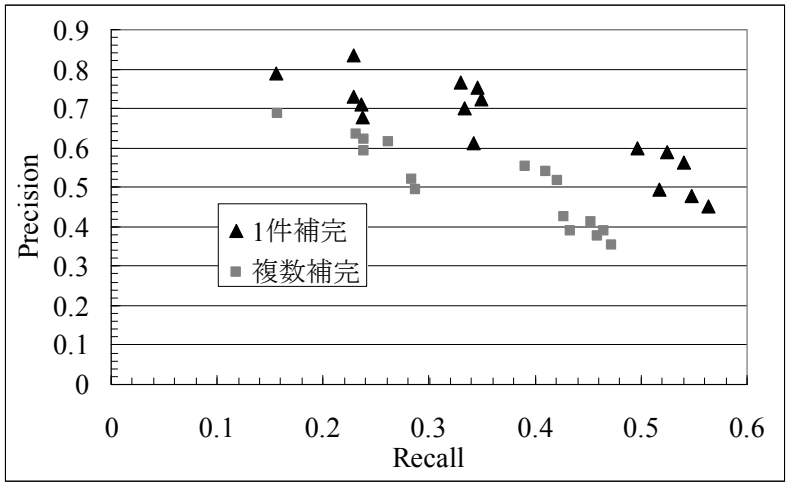

図 7　助詞落ち箇所の補完精度

助詞落ち補完精度の性能を図 7 に示す。(1)の助詞 1 つを推定する場合については、F 値が最大となるのは再現率 52.5%適合率 58.9%のときであり、再現率 22.9%のとき、最大の適合率 83.3%で助詞を補完することが可能であった。(2)の文法的に補完可能な全ての助詞を対象とした場合、F 値が最大となるのは再現率 40.1%、適合率 54.0%のときであり、再現率 15.7%のとき、最大の適合率 68.8%で助詞を補完できた。

5.4 係り受け解析精度の評価

最後に、提案手法によって助詞を補完することによる係り受け解析精度の向上について評価する。提案手法によって、係り受け解析結果が変化した文を次の 3 種類に分類する。(1)向上：補完前は不正解だった係り受け関係が補完後に正解となった場合、(2)悪化：補完前は正解だった係り受け関係が補完後に不正解となった場合、(3)不変：補完前は正解だった係り受け関係が補完後も正解である、または補完前は不正解だった係り受け関係が補完後も不正解である場合。実際に提案手法によって係り受け解析結果が変化した例を図 8 に示す。(1)の向上の例では、補完前の文を係り受け解析すると、文全体が 1 つの文節となっていたが、助詞落ち 2 箇所に対して、それぞれ「の」と「は」を補完することで正しく文節を分割することができた。(2)の悪化の例では、補完前は正しく取得できていた「限度知らず」の文節を助詞「を」を補完することで分割してしまっている。(3)の不変の例では、補完前の文は助詞落ちを含むが、係り受け関係は正しく取得できていたため、助詞「は」を補完しても係り受け解析結果に改善は見られなかった。

(1)「向上」の例
補完前：　サッカー企画面白い。
係り受け解析：　サッカー企画面白い。
補完後：　サッカーの企画は面白い。
係り受け解析：　サッカーの-D
→　企画は-D
→　面白い。
(2)「悪化」の例
補完前：一人呑みだと気楽だけど限度知らずになるからなぁ
係り受け解析：　一人-D
呑みだと---D
気楽だけど限度知らずに-D
なるからなぁ
補完後：一人呑みだと気楽だけど限度を知らずになるからなぁ
係り受け解析：　一人-D
呑みだと-----D
→　気楽だけど限度を-D
知らずに-D
なるからなぁ
(3)「不変」の例
補完前：　この荷物至極軽そうだな
係り受け解析：　この-D
荷物---D
至極-D
軽そうだな
補完後：　この荷物は至極軽そうだな
係り受け解析：　この-D
→　荷物は---D
至極-D
軽そうだな

図 8　助詞落ち箇所の補完による係り受け解析の変化

図 7 で示した提案手法において助詞を 1 つ補完する場合の再現率、適合率について 3 点を選択し、それぞれ助詞を補完した文から 500 文をサンプリングし、係り受け解析結果が(1)向上、(2)悪化、(3)不変であった割合を評価した。評価結果を表 3 に示す。最も高い再現率 52.4%のときは、2,666 箇所の助詞が補完され、向上率が 50.8%であることから、係り受け解析誤りは約 1,354 件解消されると考えられる。助詞落ちによる係り受け解析誤りは 3 節の分析で 2,735 箇所であったことから、その 49.5%は提案手法により解決できることが分かった。このとき、助詞を補完した文の 5.0%においては、係り受け解析結果に悪化が見られた。これらは閾値を調整し、適合率を高めることで軽減することが可能である。

表 3　係り受け解析精度の評価

補完の(再現率% , 適合率%)	向上率%	悪化率%	不変率%
(22.9, 83.3)	61.7	1.9	36.4
(34.6, 75.2)	58.4	3.2	38.4
(52.4, 58.9)	50.8	5.0	44.2

6. まとめ

本稿では、Twitter など一般ユーザの投稿した口語文書に頻繁に見られる助詞落ち表現が係り受け解析精度を低下させる原因の 1 つであることに着目し、解析対象となる口語文書における助詞落ち箇所を自動的に推定し、欠落した助詞を補完することで解析精度を向上する手法を提案した。提案手法では、新聞などの助詞落ちの少ない正規の文書から意図的に助詞落ちを発生させた文書を正例、助詞落ちを発生させていない文書を負例として識別器を学習させ、解析対象である口語文書の助詞落ち箇所を推定することで教師なし学習を行うことが可能な点が特徴である。加えて、推定した助詞落ち箇所の前後の単語をキーとして新聞文書を検索することで、適切な助詞を自動的に補完する。

性能評価実験においては、Twitter から収集した 10,000 文に対して提案手法を適用することで、助詞落ち箇所を再現率 22.9%、適合率 89.3%という高精度で補完できることを確認した。また、提案手法により助詞を補完した後に係り受け解析を行うことで、助詞落ちによる係り受け解析誤りを最大で 49.5%解消することが可能なことを確認した。

参考文献

1) 立石健二, 石黒義英, 福島俊一, “インターネットからの評判情報検索”, 情報処理学会研究報告, 自然言語処理研究報告, No.69, pp. 75‐82, 2001.
2) 遠藤崇史, 手塚太郎, 木村文則, 前田亮, “電子掲示板における用語間関係を用いたトピックと関係のない発言の除去手法”, DEIM フォーラム, B4-4, 2010.
3) 中島伸介, 稲垣陽一, 草野奉章, “高信頼性情報の提示を目指した熟知度に基づくブログランキング方式の提案”, 日本データベース学会論文誌, Vol.7, No.1, pp.257-262, 2008.
4) 藤村滋, 豊田正史, 喜連川優, “文の構造を考慮した評判抽出手法”, DEWS 6C-i8, 2005.
5) 渡邊拓也, 太田学, 片山薫, 石川博, “格文法を用いた複数文書融合手法”, 日本データベース学会論文誌 Letters, Vol.3, No.2, 2004.
6) T. Kudo, K. Yamamoto and Y. Matsumoto, “Japanese Dependency Analysis using Cascaded Chunking”, Proc. of the 6th Conference on Natural Language Learning (COLING) pp. 63-69, 2002
7) 河原大輔, 黒橋禎夫, “自動構築した大規模格フレームに基づく構文・格解析の統合的確率モデル”, 自然言語処理, Vol.14, No.4, pp.67-81, 2007.
8) 竹元義美, 福島俊一, “口語的表現を含む日本語文の形態素解析の実現と評価”, 情報処理学会自然言語処理研究会報告, pp.105-112, 1994.
9) Murawaki, Y. and Kurohashi, S.: Online Acquisition of Japanese Unknown Morphemes using Morphological Constraints, of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.429-437, 2008.
10) Mori, S. and Nagao, M.: Word extraction from corpora and its part-of-speech estimation using distributional analysis, Proc. of the 11th International Conference on Computational Linguistics (COLING), pp.1119-1122, 1996.
11) Ikeda, K., Yanagihara T., Matsumoto K. and Takishima Y., “Unsupervised Text Normalization Approach for Morphological Analysis of Blog Documents”, Proc. of the 22nd Australasian Joint Conference on Artificial Intelligence (AI), LNCS 5886 pp.401-411, 2009.
12) 内元清貴, 関根聡, 井佐原均, “最大エントロピー法に基づくモデルを用いた日本語係り受け解析”, Vol. 40, No.9, pp. 3397-3407, 情報処理学会論文誌, 2001.
13) 山本幹雄, 小林聡, 中川聖一, “音声対話文における助詞落ち・倒置の分析と解析手法”, 情報処理学会論文誌, Vol. 33, No. 11, pp. 1322‐1330, 1992.
14) 大野誠寛, 松原茂樹, 柏岡秀紀, 稲垣康善, “節の始境界検出に基づく独話文の係り受け解析”, Vol. 50, No. 2, pp. 553-562, 情報処理学会論文誌, 2009.
15) 船越孝太郎, 徳永健伸, 田中穂積, “音声対話用構文解析器の頑健性の評価”, 情報処理学会自然言語処理研究会報告, pp.35-41, 2002.
16) T. Kudo, K. Yamamoto, and Y. Matsumoto, “Applying Conditional Random Fields to Japanese Morphological Analysis,” Proc. of 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP) pp. 230–237, 2004.
(URL: http://mecab.sourceforge.net/)
17) R. Fan, P. Chen and C. Lin, “Working Set Selection Using Second Order Information for Training SVM,” Journal of Machine Learning Research, vol. 6 pp. 1889-1918, 2005.
(URL: http://www.csie.ntu.edu.tw/~cjlin/libsvm/)