非日本語母国話者の作成するシステム開発文書を対象とした助詞の誤用判定

大木環美† 大山浩美† 北内啓‡ 末永高志‡ 松本裕治†
奈良先端科学技術大学院大学† 株式会社 NTT データ技術開発本部 IT 活用推進センタ‡
{megumi-o, hiromi-o, matsu}@is.naist.jp {kitauchia, suenagatk}@nttdata.co.jp

1 はじめに

情報技術の発達に伴い，複雑な業務処理もシステム化されるようになってきた．このような状況の中，システムの要件をまとめて処理をフロー化する際には，文書を段階的に作成し各段階ごとにチェックすることでシステムの品質を保持している．具体的には，システムを作成する上で必要な事項をまとめた要件定義書からプログラムの機能を詳細に記述した設計書を作成し，プログラムの実装に着手する手順をとる．システムの品質を高めるためには要件定義書や設計書（まとめて仕様書と呼ぶ）の品質をまず高める必要がある．

システム開発の現場では，開発コスト削減のためにプログラム開発の海外発注が増加している．これに加え，日本語を対象にした設計書の作成の依頼も検討され始めている．その際，日本語を母語としない開発者の日本語能力によっては，日本語の文法や語彙の誤りによってシステムを実装する上で必要な機能の詳細や構成などの情報が誤って伝わる可能性がある．日本語を母国語とする熟練した技術者が校正を行えば，設計書の品質は高くなると考えられるが，一方で削減するはずであった開発コストも増加する．

コストを軽減するためには，執筆者が自ら仕様書の品質をチェックできるように対応する必要があり，図 1 のような自動的にチェックするツールによる支援が有効であると考えている．このような支援があれば，校正者が日本語の誤りを訂正すべき時間を軽減できるため，仕様書の内容に注力して校正を行うことができ，それに伴い設計書の品質が向上し，校正者の負担も軽減できると考えられる．

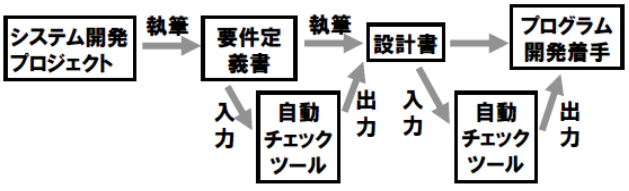

図 1: 想定するシステム開発の流れ

ここで，海外発注により執筆された仕様書 6 文書を対象に誤り傾向の調査を行った結果を報告する．対象とした 6 文書はそれぞれ，中国人の技術者が執筆した仕様書と日本語母国語話者がレビューして誤りを修正した仕様書の組で構成されている．これら修正前と修正後の文書を比較し，全体の誤りの種類を調査し，そ

表 1: 非母国語話者の執筆誤り

誤りの種類	文法	語句	内容	その他	合計
事例数	243	120	49	54	466

表 2: 文法の誤りの細分類

誤りの品目	事例数
助詞の用法	150
敬体と常体の混同	4
能動態と受動態の混同	19
時制の不一致	6
動詞と名詞句の混同	53
語順	11

の中でも頻度が高く仕様書の品質に大きく影響する文法の誤りの細分類を行った．表 1 に中国人の技術者による執筆誤りの種類を示し，表 2 に文法誤りの細分類の内容を示す．この結果から，仕様書の誤りには助詞の誤用が最も多く，全体誤りの約 3 割程度を占めることがわかった．

また，助詞は一般的に意味が抽象的かつ多義であることから日本語非母国語話者にとっては習得が容易でない（森山 [4]）ことに加え，係り受け関係や動作の対象が不明瞭になる等，仕様書の内容に大きな影響を与えることが想定される．例えば，「入力ボックス**を**挿入する」と「入力ボックス**に**挿入する」では，意味に大きな差が生じ，作成されるプログラムも大きく異なってくると考えられる．

そこで，本研究では自然言語処理を利用し，仕様書を執筆する上で留意すべき事項を執筆者が自動的にチェックできるような支援を行うために，特に助詞の誤用判定を行う技術の開発を目指す．誤用判定の対象とする助詞は，仕様書内の誤りの傾向調査において誤用されていた全 7 種類「が・を・に・の・は・により・で」とする．

2 関連研究

日本語の助詞の誤用判定を目的とした研究では，抽象化したルール，格フレーム，機械学習を用いた手法がこれまでに提案されている．

南保ら [6] は文節内の特徴を助詞と組み合わせてルール化し，帰納的学習を用いて抽出されたルール同士から抽象化したルールを新たに自動生成し，獲得されたルールに基づいて日本語の誤りの検出・校正を行うシステムを考案している．また，今枝ら [3] はルールに基づいた処理で検出できなかった事例について，格フレームをさらに適用することで検出率の改善を提案している．これらはルールベースによる手法のため，検出・校

正双方の精度の面で大変有効であるが，ルールを書くコストがかかるというデメリットがある．我々はこれらのコストもできうる限り削減できるような手法を考えている．

そこで，機械学習を用いた研究に目を向ける．Suzuki ら [2] は，最大エントロピーモデルを利用して，各文節に格助詞が必要か否か，また必要とされた場合にどの格助詞を付与すべきかをモデル化し，日本語の文構造からどの程度正確に格助詞を予測できるかを考察している．格助詞の有無を同定する正解率は 96.03%, 格助詞がある場合に格助詞をひとつ付与する正解率は 72.41%の精度を獲得している．この研究は機械翻訳における日本語文生成の準備段階として行われたものであり，学習データ，テストデータ共に新聞記事を用いている．そのため，この手法が日本語非母国語話者が執筆した文にも適用できるかは定かではない．また，Oyama ら [1] は日本語学習者による作文の校正のための格助詞の誤用検出を目的とした研究を行っており，Support Vector Machines（SVM）を用いて one vs rest による格助詞の検出手法を提案している．

3 助詞の誤用判定

仕様書の助詞の誤用を判定するにあたって，考えられるルールをすべて列挙することは多大な労力を要する上に，すべての事例を網羅することは困難である．そこで我々は機械学習を用いた手法を採用することで，より幅広い誤用に対応できる技術開発を目指す．

我々は Oyama らの提案手法を基本的に踏襲しており，さらに仕様書に特化した素性生成方式の提案を行う．

3.1 基本手法

Oyama らは，文中で使用される助詞は前後の語に依存するという考察から，正しい日本語で書かれた文書から助詞の周りに出現する語と出現しない語を学習することで，助詞の用例を誤用例と正用例に分類することを行っている．例えば (1A) で太字で示した助詞は正しい用法であるが，(1B) は誤った助詞の用法である．しかし，(1C) や (1D) では助詞「に」の用法は正しい．このように，使用すべき助詞が助詞の前に出現する主語や助詞の後に続く動詞に依存して決定されることがわかる．

(1) A 医師の指示**で**麻酔を投与する
B 医師の指示**に**麻酔を投与する
C 医師の指示**に** 従う
D 患者 **に**麻酔を投与する

助詞の誤用検出処理は図 2 のように行われ，正しい日本語のデータとして新聞記事を学習データに用い，助詞ごとに素性を作成する．図 2 では助詞「に」に対しての素性例を示している．

新聞記事で用いられている助詞はすべて正しい用法であるとすると，負例の学習データが得られない．そこで，負例の学習データには疑似負例を用いる．疑似負例では，例えば，正しい用法「が」の事例を助詞「に」の負例として扱うといったように，他の助詞で用いられている事例を負例とみなす．ここで注意すべきは，他の助詞を対象とした事例が必ずしも負例であるとは限らないという点である．例えば,「芥川賞が決定」という事例は助詞「に」を対象とした学習データでは，誤りの事例として学習されることとなるが，助詞「に」の事例（「芥川賞に決定」）として考えてもこれらの情報のみでは誤りとは考えられない．このような日本語の曖昧性の問題は大変興味深いが，本研究ではこの問題への対応は今後の課題とする．

また，日本語母国語話者が修正した仕様書の助詞の誤りのデータは大変手に入りにくく，データ数としても少ないため，本研究でも疑似負例を採用することとする．学習データから作成した素性で SVM を用いて学習を行い，テストデータの素性を分類器に適用する．図 2 の場合，助詞「に」の用法として正しい事例を正例，誤りの事例を負例と判定できれば成功である．

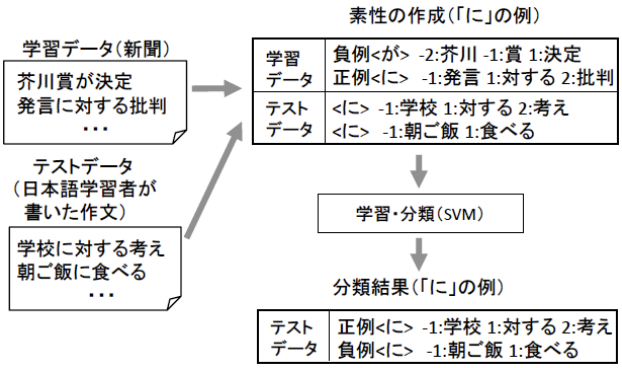

図 2: 誤用検出処理の一例

学習のための素性には，その助詞から前後 3 単語の原形と品詞を用い，助詞の係り先の動詞の情報も素性に加える．さらに Oyama らは，文内の固有表現（人名・数詞・土地・アルファベット）の統一化をはかることで素性が疎になりすぎないような操作を行っている．

Oyama らの基本手法による素性生成の処理と例を表 3(I〜IV) に示す．表 3(I) は，例「５歳で太郎は自転車に乗れるようになった」に対する形態素解析[1]，構文解析 [5] の結果を示す．“[]” 内には形態素解析で得られた文節，“_” は形態素の区切りをそれぞれ示し，“()” には構文解析により得られた係り先の番号を記した．文節番号は始めを 0 として数え，例えば，“(3)” は文節 [乗れるようになった] を示す．表 3(II) では，固有表現である数詞「５」と人名「太郎」をそれぞれ「Number」,「Person」といった特定の記号に置き換えている．表 3(III) で，助詞「に」を対象とした場合に生成される素性列を示し，負の数字は単語が対象とする助詞からいくつ前の位置にあるか，正の数字は対象とする助詞からいくつ後の位置にあるかを表している．最後に表 3(IV) で，助詞「に」の係り先である文節「乗れるようになった」内の動詞の原形と品詞および機能表現を D から始まる素性として示している．

3.2 素性生成の改善

仕様書に特化した技術を開発するために，我々は前節で示した基本手法の素性生成方式に対して，以下に示すような４つの改善案を考案した．

3.3 素性の形式の変更

基本手法の素性生成では単語と品詞の組み合わせを素性として扱っており，素性が疎になる可能性がある．

[1] http://chasen.org/ taku/software/mecab/

表 3: 基本手法（I・II・III・IV）および改善案（i・ii・iii）による素性生成例

処理	例（5歳で太郎は自転車**に**乗れるようになった）				
I. 文の形態素解析・構文解析	[5_歳_で (3)][太郎_は (3)][自転車_に (3)][乗れる_よう_に_なっ_た]				
II. 固有表現 (人名・数詞・地名・アルファベット) の統一化	[Number_歳_で (3)][Person_は (3)][自転車_に (3)][乗れる_よう_に_なっ_た]				
III. 対象とする助詞の前後 3 単語を取得	-3:Person:名詞-固有名詞-人名 -2:は:助詞-係助詞 -1:自転車:名詞-一般	に	1:乗れる:動詞-自立 2:よう:名詞-非自立-助動詞語幹 3:に:助詞-格助詞-一般		
IV. 対象とする助詞の係り先内の動詞の単語と機能表現の取得	-3:Person:名詞-固有名詞-人名 -2:は:助詞-係助詞 -1:自転車:名詞-一般	に	1:乗れる:動詞-自立 2:よう:名詞-非自立-助動詞語幹 3:に:助詞-格助詞-一般	D:乗れる:動詞-自立 D:なる:動詞-自立 D:た:機能表現	
i. 素性の形式の変更	-3:Person -3:名詞-固有名詞-人名 -2:は -2:助詞-係助詞 -1:自転車 -1:名詞-一般	に	-1:乗れる -1:動詞-自立 2:よう 2:名詞-非自立-助動詞語幹 3:に 3:助詞-格助詞-一般	D:乗れる D:動詞-自立 D:なる D:動詞-自立 D:た D:機能表現	
ii(a). 同じ係り先の素性を持つ助詞の原形と品詞の素性の追加	-3:Person:名詞-固有名詞-人名 -2:は:助詞-係助詞 -1:自転車:名詞-一般	に	1:乗れる:動詞-自立 2:よう:名詞-非自立-助動詞語幹 3:に:助詞-格助詞-一般	D:乗れる:動詞-自立 D:なる:動詞-自立 D:た:機能表現	hasD:で:助詞-格助詞-一般 hasD:は:助詞-係助詞
ii(b). 対象とする助詞の前後の単語の活用形の素性の追加	-3:Person:名詞-固有名詞-人名 -2:は:助詞-係助詞 -1:自転車:名詞-一般	に	1:乗れる:動詞-自立:基本形 2:よう:名詞-非自立-助動詞語幹 3:に:助詞-格助詞-一般	D:乗れる:動詞-自立 D:なる:動詞-自立 D:た:機能表現	
iii. 助詞の存在しない学習事例の追加	-3:は:助詞-係助詞 -2:自転車:名詞-一般 -1:に:助詞-格助詞-一般	**NONE**	1:乗れる:動詞-自立 2:よう:名詞-非自立-助動詞語幹 3:に:助詞-格助詞-一般	D:乗れる:動詞-自立 D:なる:動詞-自立 D:た:機能表現	

そこで我々は，表 3(i) に示すように単語と品詞を独立した素性とし，２次の多項式カーネルを学習器に利用することで，これらの組み合わせが自動的に考慮されるようにした．

i 新たな素性の追加

a. 同じ係り先の素性を持つ助詞の原形と品詞

基本手法では，前後 3 単語と係り先の文節に含まれない単語は素性の情報として扱われていなかった．しかし，主語・目的語を明確に示す助詞は他の助詞との関連も深く，対象としている助詞の誤用を判定する上で有用な情報であると考えられる．例えば，「表**を**挿入する」は文脈によっては「表**に**挿入する」とも矛盾なく書くことができ，一意に助詞を決定することは困難である．しかし，「設計書**に** 表**を**挿入する」という文の場合には，助詞「に」を持つ目的語が既に存在しているため「表**に**挿入する」は誤用であるということがわかる．

表 3(iia) では hasD から始まる素性が同じ係り先の素性を示しており，前後 3 単語に出現しない助詞「で」の情報を得ることができる．

b. 対象とする助詞の前後の単語の活用形

基本手法では，単語の活用形が素性に加えられていないために，動詞の基本形と動詞の連用形が同じ素性として扱われてしまう等の問題点があった．例えば，「データがない**の**場合」では助詞「の」は誤用であるが，「データがなし**の**場合」は正用である．ここで問題となるのは，双方の事例で助詞「の」を対象として素性を生成した場合に，同じ素性列となってしまうことである．このような動詞の連体化は仕様書では多く見られる．そこで，我々は対象とする助詞の前後の単語の活用形を素性に追加することでこの問題に対処した．表 3(iib) では，助詞「に」の前後の単語「自転車」と「乗れる」に対して活用形がある場合に素性を追加する．

3.4 助詞の存在しない学習事例の追加

基本手法では，誤用を判定する助詞のみを対象として学習データの事例を生成しているが，実際の助詞の誤用には本来助詞が不要な箇所に助詞を誤って挿入してしまうような事例が存在している．そこで，助詞が存在しない事例を学習することでこの問題に対処する．

表 5: データの内訳

助詞	対象助詞		助詞なし	合計
	正用例	誤用例		
学習データ	203,807	-	428,876	632,683
テストデータ	4,538	90	-	4,628

名詞間などでは助詞の省略が可能になる場合もあることから，助詞が存在しない箇所すべてを対象とするのは，助詞の誤用を判定する上で悪影響を及ぼす懸念がある．

そこで，対象とする助詞の前後の単語の品詞をペアとしてすべて抽出し，その品詞のペアが前後に出現しない事例のみを対象とした．例えば，表 3 の例では助詞の前後の単語の品詞のペアとして，“名詞-数，名詞-固有名詞-人名” や “名詞-一般，動詞-自立” 等が抽出され，隣り合う単語の品詞がこれらの品詞のペアに相当しない場合に，助詞がない事例の素性列が生成される．助詞がない事例では助詞の箇所を「NONE」として記述することとする．

4 誤用判定の評価実験

4.1 実験設定

我々の提案した素性生成方式の改善が，仕様書の誤用判定を行う上で有効であるかを検証する．本研究は助詞の誤用を判定することに焦点を置くが，実際にツールとして使用する場合，正しい用法で用いた助詞を誤っていると判定してしまうことは，技術者のツールに対する信頼度を下げることにつながるため，正用例に対する判定性能も同時に評価することとする．学習データは複数の分野の仕様書から抽出した事例を用い，テストデータは傾向調査で使用した仕様書内の誤用例と正用例を用いた．データの詳細を表 5 に示す．実験は基本手法を方式 1 とし，素性生成の方式を改善した各手法の性能を方式２〜５，改善案すべてを行った素性生成方式を方式 6 としてそれぞれの結果を比較した．

4.2 結果と考察

結果を表 4 に示す．それぞれの実験はチェックマークのついた素性生成方式を用いており，素性生成方式の記号は表 3 と対応する．また，正用例・誤用例の性能の向上を矢印で示した．すべての改善案を取り入れた方式 6 が最も高い性能を示した．誤用例の precision が

表 4: 実験結果

	素性生成方式					precision(%)		recall(%)		性能	
	基本手法	i	ii(a)	ii(b)	iii	正用例	誤用例	正用例	誤用例	正用例	誤用例
方式 1	√					99.4(3,582/3,603)	6.7(69/1,025)	78.9(3,582/4,538)	76.7(69/90)	-	-
方式 2	√	√				99.3(3,798/3,823)	8.1(65/805)	83.7(3,798/4,538)	72.2(65/90)	↗	↘
方式 3	√		√			99.4(3,772/3,795)	8.0(67/833)	83.1(3,772/4,538)	74.4(70/90)	↗	↘
方式 4	√			√		99.4(3,563/3,584)	6.6(69/1,044)	78.5(3,563/4,538)	76.7(69/90)	↘	→
方式 5	√				√	99.5(3,435/3,552)	6.2(73/1,176)	75.7(3,435/4,538)	**81.1**(73/90)	↘	↗
方式 6	√	√	√	√	√	**99.5**(3,873/3,893)	**9.5**(70/735)	**85.3**(3,873/4,538)	77.8(70/90)	↗	↗

全体的に低い値となっているが，正用例のデータ数に比べ，誤用例のデータ数が少ないため precision の低下が著しく見られる．しかし，本実験では実際の仕様書の正誤用例を用いているため，実際にツール化した場合の性能と結果は同様の傾向を示すと考えられる．このことからも，全体の性能を向上するためには正用例の判定性能もより高く保つ必要がある．以下に方式２〜５における考察を示す．

方式２

素性が疎になる状態が改善されたため，正用例の判定性能が向上したと考えられるが，一方で，素性の組み合わせを考えることにより，係り先が存在しないような不自然な文の事例を誤って判定する傾向が見られた．仕様書は主にエクセルを用いて執筆され，一文が複数のセルにまたがり，一つのセル内で文が完結していない事例が多々存在するため，このような事例が多く含まれていると考えられる．

方式３

助詞「の」の正用例の判定性能の向上が顕著に見られる．これは助詞「の」が他の助詞と同じ係り先を持ちにくい性質があると考えると，同じ係り先を持つ助詞の素性が存在しないことが助詞「の」を判定する上で有用な情報となったと考えられる．また，誤判定の事例はどれも文内の他の助詞が誤っていることが見受けられた．このことから，素性自体は誤用例の適用にも効果があるが，文内で複数の助詞が誤っている場合に問題があると考えられる．同じ係り先を持つ助詞を素性として加えない場合には，誤用例として正しく判定できていることから，段階的な誤用例の検出や同時に文内のすべての助詞の組み合わせを判定する等の工夫が必要である．

方式４

正用例の判定性能が低下したが，これは単語と品詞に活用形を加えた素性としたため，素性が疎になってしまったことが一因であると考えられ，方式２を用いて単語，品詞，活用形をそれぞれ別の素性として２次カーネルで学習した場合には，わずかながら性能の改善が見られている．

方式５

この結果から，助詞の存在しない事例は誤用例を判定する性能の向上に大きく貢献することがわかった．詳しくエラー分析を行うと，本来助詞「の」が存在しない箇所に誤って記述してしまった事例を誤用として判定することができている．これは素性を追加した狙い通りの結果が得られたと言える．しかし，助詞の存在しない事例として正用例の助詞「の」を誤って誤用例として判定してしまっている事例も多い．連体化を示す助詞「の」については省略が可能である場合が多く，一概には誤った判定とは言えないことから，実用に関しては問題ないと考えられる．

5 おわりに

海外発注で執筆された仕様書の品質を保持するために，仕様書の執筆において留意すべき点を執筆者が自動的にチェックできるツールを作成することを目標とし，本研究は，特に非日本語母国語話者による仕様書内の誤りの中で最も多い助詞の誤り検出の技術開発を目的とする．本研究では Oyama らの機械学習による誤用判定手法における素性生成方式の改善を行い，我々の提案する方式の仕様書に対する誤用判定性能の評価を行った．我々の提案した改善案は，仕様書の助詞の誤用を判定する上で有用であるが，正用例の判定性能をより改善する必要があることがわかった．本稿では素性の追加および変形に目を向けた改善案を提案したが，今後は学習に不必要な素性の削除によって，素性の質の改善を図り，実際に現場で使用できるツールとして発展させていくことを考えている．

参考文献

[1] Hiromi Oyama, Yuji Matsumoto, Masayuki Asahara, and Kosuke Sakata. Construction of an error information tagged corpus of japanese language learners and automatic error detection. In *Proceedings of the Computer Assisted Language Instruction Consortium*, 2008.

[2] Hisami Suzuki and Kristina Toutanova. Learning to predict case markers in japanese. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 1049–1056, Sydney, Australia, July 2006. Association for Computational Linguistics.

[3] 今枝恒治, 河合敦夫, 石川祐司, 永田亮, 桝井文人. 日本語学習者の作文における格助詞の誤り検出と訂正. 情報処理学会報告 2003-CE-68.

[4] 森山新. 認知言語学から見た日本語格助詞の意味構造と習得. シリーズ言語学と言語教育 16. ひつじ書房.

[5] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.

[6] 南保亮太, 乙武北斗, 荒木健治. 文節内の特徴を用いた日本語助詞誤りの自動検出・校正 (語学学習支援・自動校正). 情報処理学会研究報告 2007-NL-181.