機械学習による日本語格助詞の予測

Hisami Suzuki　　Kristina Toutanova
Microsoft Research
One Microsoft Way, Redmond WA 98052 USA
{hisamis, kristout}@microsoft.com

概要

日本語の格助詞は、述語の補足語である名詞句が述語に対してどのような関係にあるかを示す役割を持つが、その使い分けは日本語学習者にとっても、また日本語文の自動生成においても難しいものである。本稿では、機械翻訳における日本語文生成の準備段階として、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のかを考察する。本稿で提案する予測手法は、英語における名詞句の意味役割付与 (semantic role labeling) タスクに着想を得たもので、各々の文節に格助詞が必要か否か、必要と判断された場合どの格助詞を付与するかを最大エントロピーモデルを使ってモデル化している。京大コーパスを用いて実験を行った結果、総文節中の82%に正しく格助詞が付与された。

1　はじめに

自然言語処理における文生成の役割は、機械翻訳技術の進展とともに、今後さらに重要となってくることが予想される。しかしながら、現時点での英日機械翻訳における日本語文生成には難題がいくつかあり、その中でもとわけ難しいのが日本語格助詞の生成である。格助詞の使い分けは日本語学習者にとっても非常に難しいものとされているが、これは格助詞が翻訳における原文、あるいは日本語学習者における母国語に対応する語がないことが多く、また、格助詞とそれが表している述語との関係が非常に複雑であることなどが理由であると考えられる。内容語と文の構造から格助詞を適切に予測するモデルを構築することができれば、それは機械翻訳や日本語生成一般に有益であると思われる。

言うまでもなく、格助詞の生成は内容語の選択に加えて、文の意味や発話の意図が大きく関与してくる。たとえば、「私」「パスタ」「食べました」という内容語が与えられたとき、これらの語の間に成立する関係は複数考えられる。「私□パスタ□食べました」という語の並びを考えると、一つめの□には「は」「が」「と」などが、二つめには「を」「は」「から」などが適当であるが、一つめには「を」「に」「で」、二つめには「が」「に」「で」などは不適当である。このように、どのような助詞を生成するのかをひとつの正解に照らして正確に予測することは困難であるが、少なくとも明らかに不適格なものを除外することは可能である。機械翻訳などの応用技術では、意味的な制約は翻訳モデルから得ることができるので、格助詞の適格性をこのようなモデルを使って相対的に評価することができれば、翻訳の質の向上につなげることができる。

本稿ではこのような観点から、機械翻訳における日本語文生成の準備段階として、日本語の文構造からどの程度的確に格助詞を予測できるのかを考察する。本稿で提案する予測手法は、英語における名詞句の意味役割付与 (semantic role labeling) タスク[4,5,7]に着想を得ているが、これはPropBank[6]でタグづけされている20の意味役割を予測するタスクであり、機械学習による分類器を用いて、非常に高い予測精度を得ている。本稿では、まず2節で日本語の格助詞の予測タスクを定義し、3節で提案手法について述べる。日本語の格助詞予測タスクは、英語の意味役割付与タスクに比べてはるかに曖昧性が大きく、難度の高いタスクであるが、4節で報告するように、京大コーパスにおいて総文節中の82%に正しく格助詞が付与される、という結果を得た。

2　格助詞予測タスクの定義

2.1　日本語の助詞について

本稿の格助詞予測タスクでは、日本語の名詞句につく助詞の一部をその対象とした。対象とする助詞を選択するにあたり参考にしたのが、以下3種の助詞である(Cf. [2,3])。

格助詞. 格助詞は補足語が述語に対してどのような関係にあるかを表すもので、補足語と述語の両方によって選択される。この特徴が、機械翻訳などにおける格助詞の生成を困難にしている。また、格助詞とそれがあらわす意味の結びつきは大変複雑であり、「東京に住む」「東京に行く」のように、ひとつの格助詞が複数の意味役割を担うこともあれば、「東京に住む」「東京で会う」のように、ひとつの意味役割が複数の格助詞で表されることもある。本稿で扱う格助詞予測タスクでは、「が、を、の[1]、に、から、と、で、へ、まで、より」の格助詞すべてをその対象とした。

[1] 「の」は通常、接続助詞として分類されているが、機械翻訳においては「の」の生成も重要なこと、また「の」は関係節内では格助詞として機能することから、タスクの対象に含めた。

接続助詞. 名詞や名詞句をつなぐ接続助詞は英語の "and" "or" などに相当し、文中の情報のみから予測するのは困難として、本タスクの対象外とした。

取り立て助詞. 取り立て助詞は、「パスタしか食べなかった」「パスタも食べた」の「しか」や「も」のように、ある文脈や背景に対して補足語を取りあげる役割をする。取り立て助詞もその性質から、文中の情報のみからは予測不可能なので、本タスクの対象からはずした。ただし、「は」は、取り立て助詞のひとつとして分類されることもあるが以下の理由から本タスクの対象に含めた。

- 「は」の主要な働きは主題を提示することであり、この点、ほかの取り立て助詞と異なり、「提題助詞」と分類されることもある。主題は文の構造上からある程度予測可能である。
- 「は」は、日本語の助詞の中で「の」「を」についで頻度が高く[2]、したがって「は」を適切に生成することは日本語文生成にとって重要である。
- 「は」は、英語文に相当する語句がないことから、統計的な手法に基づいた英日翻訳システムで生成することが大変困難である。

以上、10の格助詞と「は」に加えて、格助詞と「は」から成り立っている「には、からは、とは、では、へは、までは、よりは」の7つもそれぞれ格助詞とみなし、合計で18の助詞を予測タスクの対象とした。

2.2 タスクの定義

本稿の実験の対象とした格助詞予測タスクでは、まず文節間の係り受け情報つきの日本語文が与えられているものとする。ここでは京大コーパス[1]を使用した。このコーパスから得られる情報を用いて、各文節に格助詞が付与されているか、されていると判断された場合には18のうちのどの格助詞が適当かの分類を学習する。評価時には、文節から格助詞の情報を取り除いて、それがどの程度正確に予測されるかを評価する。すなわち各文節を、18の格助詞と格助詞なし(NONE)の計19のクラスに分類するタスクとなる。

3 格助詞予測の提案手法

3.1 予測モデル

提案手法では、英語の意味役割付与タスクに倣い[5,7]、まず格助詞予測タスクを、格助詞同定 (identification) と分類 (classification)のふたつのサブタスクに分けて考える。格助詞同定のサブタスクでは、各文節を、HasCase (格助詞あり)か、NONE（格助詞なし）の2種に分類する。格助詞分類のサブタスクでは、同定タスクにより「格助詞あり」と判定された文節に対して、18の格助詞の中からひとつを付与する。ここで、cをNONEを含めた格助詞分類のクラス、bを文節とすると、同定モデルによる確率は$P_{ID}(c|b)$、分類モデルによる確率は$P_{CLS}(c|b)$と書ける。同定と分類を含めた予測タスク全体の確率は次のように表すことができる。

$$P_{CaseAssign}(NONE|b) = P_{ID}(NONE|b)$$

$$P_{CaseAssign}(l|b) = P_{ID}(\text{HasCase}\ |b)^* P_{CLS}(l|b)$$

ここで、l は18の格助詞のいずれかである。

このように予測タスクをサブタスクに分けることにはふたつの利点がある。ひとつは、サブタスクごとに固有の素性を使用することが可能になる点である。ただし本稿の実験では、ふたつのサブタスクに同じ素性を使用した。もうひとつの利点は学習効率の向上である。これは、サブタスク化によって、分類モデルは学習の際、格助詞つきの文節のみをサンプルとして使用すればよくなるからである。

分類モデルの学習には、さまざまな手法が適用できるが、ここでは最大エントロピー法を使用した。最大エントロピー法は確率分布を出力するので、ふたつのサブモデルを組み合わせたり、さらに機械翻訳システムの一部として組み込むのに都合がよい。

3.2 使用した素性

実験に使用した素性を表1に示す。素性は格助詞の予測がなされている文節の素性、その係り先の文節の素性、さらにそのふたつの文節間の関係に関する素性、の3種に分かれる。これらの基本素性に加えて、素性の組を20個使用した。そのうちの3つを表1に示している。これらの素性は人手で作られたものであり、組み合わせもまだその一部しか使っていないことから、さらに改良の余地があると思われる。また、素性は学習の際にはすべて二値の素性として与えられている。すなわち、表1で二値以上の値をとる素性は、その素性名と値の組を、固有の素性として扱っている。モデルは、正規分布を使って正則化した。

4 実験結果と考察

4.1 実験データ

この実験には、京大コーパス (Version 3.0) [1]を次のように分割して使用した。

	コーパス	総文数	総文節数
学習データ	記事1月1日、3日–11日分 社説1月–8月分	24,263	234,474
ディベロップメントデータ	記事1月12,13日分 社説9月分	4,833	47,580
テストデータ	記事1月14–17日分 社説10–12月分	9,287	89,982

[2] 京大コーパスでは、「の」が全助詞のうちの20.6%,「を」が13.5%、「は」が13.2%を占める。

Basic features for phrases (self, parent)
HeadPOS
HeadNounSubPos: time, formal nouns, adverbial
HeadLemma
LastWordLemma (excluding case markers)
LastWordInfl (excluding case markers)
IsFiniteClause
IsDateExpression
IsNumberExpression
HasPredicateNominal
HasNominalizer
HasPunctuation: comma, period
HasFiniteClausalModifier
RelativePosition: sole, first, mid, last
NSiblings (number of siblings)
Position (absolute position among siblings)
Voice: pass, caus, passcaus
Negation
Basic features for phrase relations (parent-child pair)
DependencyType: D,P,A,I
Distance: linear distance in bunsetsu, 1, 2-5, >6
Subcat: POS tag of parent + POS tag of all children + indication for current
Combined features (selected)
HeadPOS + HeadLemma
ParentLemma + HeadLemma
Position + NSiblings

表 1: 基本素性と素性の組

本実験ではディベロップメントデータは素性の抽出と評価にのみ使用した。また、学習データは評価にも使用した。

4.2 実験結果と考察

格助詞予測の実験の結果を表2にまとめる。まず、上から3行は提案手法による正解率である。上から順に、すべての文節に対しての同定タスク (Identification) の正解率、格助詞のある文節に対しての分類タスク (Classification) の正解率、そしてすべての文節に対しての格助詞予測モデル全体(Ident+Classif)の正解率である。テストデータにおける予測モデル全体の正解率は82.27%であった。

表2の下3行は、ベースラインモデルの正解率を示している。最も単純なベースラインは、18の格助詞にNONEを加えた19のクラスのうち、もっとも頻度の高い分類(=NONE)を常に選択するもので、この方法の正解率は47.41%であった。さらに現実的なベースラインとして、word trigram をつかった言語モデルとも比較した。Nグラムに基づく言語モデルは、タグつきコーパスを必要とせず、大規模なデータからの構築が可能なので、このようなタスクにも十分有益であると考えられる。ここでは、4.1節の学習データ(24,263文)を使ったモデル (KCLM) と、単語分割情報のみ与えられた大規模な新聞データ(826,373文)を使ったモデル(BigCLM)の2種の言語モデルを構築した。BigCLMの学習では、bigramと trigramはコーパスに最低5回出現したもののみを使用した。これらの言語モデルを、提案手法の解析と同じく、各々の文節に格助詞またはNONEのラベルを付与するために使用した。その際、最適な格助詞列は動的計画法を使って求めた。

タスク	学習データ	テストデータ
Identification (subtask)	98.41	96.03
Classification (subtask)	88.66	72.41
Ident+Classif	92.68	82.27
baseline (frequency)	47.41	47.41
baseline (KCLM)	93.92	66.98
baseline (BigCLM)	—	76.12

表 2: 格助詞予測モデルの正解率 (%)

これらの言語モデルを使ったベースラインの結果であるが、まずKCLMは学習データでは高い正解率(93.92%)を示したが、テストデータでの正解率は、提案手法の82.27%に遠く及ばず、66.98%にとどまった。このことはNグラムモデルでの学習がうまく一般化されないことを示している。BigCLMでは、テストデータにおける正解率が76.12%まで向上し、学習データの量の重要性が示されている。ただし、このベースラインと比較しても、提案手法の不正解率はさらに25%改善されている。。また、今回の実験には取り入れなかったが、将来的に、Nグラム言語モデルを素性として提案手法に組み込むことも可能である。

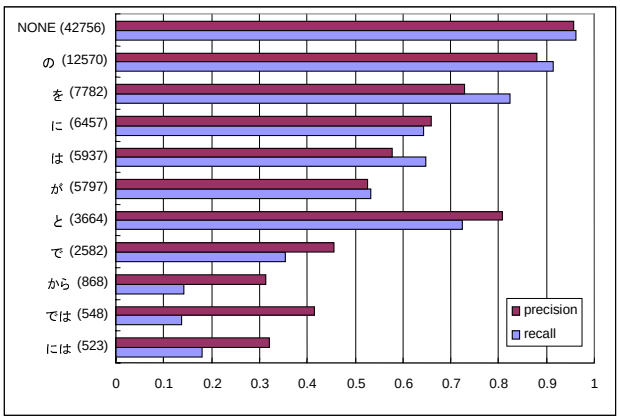

図 1: 格助詞ごとの適合率 (precision) と再現率 (recall). (ともに単位は%. カッコ内はテストデータにおける頻度)

図1に、頻度の高い格助詞(テストデータでの頻度が500以上)の適合率 (precision) と再現率 (recall)を示す。「適合率」はシステムが出力した格助詞のうち正解だったものの率、「再現率」は正解データに含まれる格助詞のうち、システムが出力したものの率である。「と」を除いて、全体的に頻度と適合率・再現率の間に正の相関が見られる。NONEと「の」の正解率が高いのは、この2つ

	phrase	predicate	sentence
1-best	82.27	78.34	21.18
2-best	88.73	86.20	37.20
5-best	94.19	92.66	60.05
10-best	96.69	95.71	74.48
20-best	98.09	97.49	84.41

表 3: 基本素性と素性の組

の分類が比較的曖昧性が少ないことも影響していると思われる。これに対し、「に」「は」「が」「で」などは正解以外の格と文法的にも意味的にも交換可能な場合も多く、それだけ正解率も低い。特に「では」「には」などの、格助詞と「は」の組み合わせは、「で」「に」などの格助詞と交換可能性が高く、ディベロップメントデータから[正解＝に: システム解=には]と[正解＝には: システム解=に]のケースを各10個ずつ、計20個抜き出して調べたところ、交換不可能なケースは2つのみであった。このような助詞と助詞＋「は」の組を同じクラスに属するものとして評価しなおしたときの正解率は、格助詞分類のサブタスクで73.2%、予測タスク全体で82.6%と、多少の向上が見られた。

最後に、本稿で提案されたモデルを機械翻訳システムの一部として使用することを念頭におき、格助詞予測モデルを使ってNベスト解を出力した結果を表3に示す。ただし、表3における「Nベスト解」は、各文節についてのNベストではなく、ひとつの述語にかかる文節をひとまとめに扱ったときのNベストである。表3の「述語正解率」は、述語にかかる文節の格助詞がすべて正しく付与された率、「文正解率」は、文中の文節のすべてに格が正しく付与された率である。Nの値が大きくなるにつれ、正解率も上がっていることから、機械翻訳システムなどの一部として使用する際には、Nベスト解の使用が適切であると思われる。

5 関連研究

日本語の助詞や付属語の生成を扱った研究としては[8]がある。[8]では、文節の主辞となりうる自立語から、格助詞に限らず、付属語一般を生成するモデルを提案している。モデルは複数のNグラム言語モデルからなり、モデル構築には文節の係り受け構造と、表層の単語列を使用している。[8]の評価方法は、本稿とは異なり、自立語の2つ組みと3つ組みから付属語表現を生成し、人手で評価しているため、本稿との比較は困難である。また、手法も本稿で提案したモデルとは異なっているが、[8]で提案されているNグラムモデルは、本稿で提案されたモデルにおいても、素性として使用が可能である。

6 おわりに

本稿では、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のかを、機械学習を使い、京大コーパスを用いて実験した。現時点での実験結果をもとに、さまざまな応用実験や改良実験が考えられるが、とりわけ重要なのは提案されたタスクと手法を、機械翻訳システムの一部として使用し、評価することである。さらに、今回使用したモデルは、文節の格付与の確率をほかの文節の格を考慮せずに予測する局所的なモデルであるが、英語の意味関係付与タスクでは、ほかのノードに付与された意味関係を取り入れた大局的なモデルを局所モデルと一緒に使用することによって、正解率が大幅に改善されている[7]。日本語の文節に与えられる格もまた、文全体の中で決まってくることから、格助詞予測においても大局的なモデルの使用は有益であると思われる。

参考文献

[1] 黒橋禎夫、長尾真. 1997. 京都大学テキストコーパスプロジェクト. 言語処理学会第3回年次大会誌. pp.115-118.

[2] 寺村秀夫. 1991. 日本語のシンタクスと意味 第III巻. くろしお出版.

[3] 益岡隆志、田窪行則. 1992. 基礎日本語文法—改訂版—. くろしお出版.

[4] Carreras, Xavier and Lluís Màrquez. 2004. Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling. In *Proceedings of CoNLL-2004*.

[5] Gildea, Daniel and Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. In *Computational Linguistics* 28(3): 245-288.

[6] Palmer, Martha, Dan Gildea and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. In *Computational Linguistics* 31(1).

[7] Toutanova, K., Aria Haghighi and Christopher D. Manning. 2005. Joint Learning Improves Semantic Role Labeling. In *Proceeding of ACL*, pp.589-596.

[8] Uchimoto, Kiyotaka, Satoshi Sekine and Hitoshi Isahara. 2002. Text Generation from Keywords. In *Proceedings of COLING*.