機械学習を用いた複合動詞の多義性解消

内山清子†‡
‡ 慶應義塾大学大学院
政策・メディア研究科
〒252-8520 神奈川県藤沢市遠藤 5322
<kiyoko@sfc.keio.ac.jp>

Timothy Baldwin†
† Center for the Study of
Language and Information (CSLI)
210 Panama Street, Stanford,
CA 94305-4115, USA
<tbaldwin@csli.stanford.edu>

概 要

本研究は，複合動詞の多義性を解消するために，新聞記事コーパスから抽出した前項動詞と後項動詞の共起情報を学習させることにより，新規の複合動詞の意味を特定する手法を提案する．本手法を用いて学習を行った結果，約 86%の F 値を得て，手法の優位性を確認した．

1 はじめに

複合動詞は連続動作だけでなく，動作の状態や時間の変化などを表現するために，様々な前項動詞と後項動詞を結合して生成することができる．本研究では，動詞 2 語から成り，語構成要素が単独で用いられる時の意味から，複合動詞の意味を推測できる「押し上げる（上に押して上げる）」などの，様々な前項動詞と後項動詞が結合して生成される生産的な複合動詞を分析する．

生産的な複合動詞の中には，同じ後項動詞を用いて複数の意味を持つものがある．たとえば，後項動詞「上げる」が，複合動詞の「ボールを蹴り上げる」と「野菜を茹で上げる」において，前者は「上方向に上げる」という空間移動の意味を持つ一方，後者は「茹でることを終える」という時制的意味を表している．このように複数の意味を持つ後項動詞を多義的後項動詞と呼び，多義的後項動詞から成る複合動詞を本研究の対象とする．

後項動詞の多義性は前項動詞の意味的特徴に関連があり，身体動作「蹴る」「投げる」「押す」などが前項動詞の場合，後項動詞「上げる」は空間的移動を表し，「茹でる」「蒸す」「焼く」などの調理の動作では，後項動詞が時制的意味を表す傾向にある．こうした前項動詞の意味的特徴と後項動詞との組み合わせをルール化して多義解消を試みた研究 [1] を行い，作成したルールを辞書データにより評価した結果が約 87%と高かったが，前項動詞の意味属性を自動的に決定することが難しく，課題として残った．

そこで，本研究では前項動詞の意味属性を用いずに，新規の複合動詞の意味を特定する手法を提案する．以下第 2 章では，研究の背景を説明し，第 3 章では複合動詞の多義性及びその種類について論じ，第 4 章では機械学習を用いた多義性解消実験の詳細を述べ，最後に考察を行う．

2 研究の背景

日本語の複合動詞に関する研究は，言語学や自然言語処理の分野において数多く議論されている．言語学の分野では，形態・語彙論的分析により語構成要素の文法的違いに基づいて，統語的複合動詞と語彙的複合動詞に分類した研究 [2]，複合動詞を構成する主な後項動詞の意味用法を分析した研究 [3] が行われている．本研究で扱う複合動詞の多義性は，統語的複合動詞と語彙的複合動詞の違いや語彙的複合動詞内における意味の違いを含んでいる．従来の研究における統語的複合動詞と語彙的複合動詞を分類する基準は，言語的直感に依存した部分が大きく，分類の自動化に応用することが難しい．本研究が目的とする多義性解消の手法は，統語的および語彙的複合動詞の識別に対しても有効な手段であると考える．

また，自然言語処理の分野では，共起する名詞と複合動詞をセットにして辞書に一括登録する手法 [4] が提案されている．複合動詞を辞書に 1 つの動詞として一括登録しておく方法は，便宜的で有効な手段であるが，新規の複合動詞を処理できず，生産性の高い複合動詞を処理することが難しい．そこで，複合動詞の個々の語構成要素を単位として，その組み合わせにより意味を特定する

In Proceedings of the 10th Annual Meeting of the Association of Natural Language Processing (Japan), Tokyo, Japan (in Japanese).

方法を取る．本研究では，前項動詞と後項動詞の組み合わせにより生成される複合動詞の意味の特定に，機械学習のモデルの採用を試みた．

3　複合動詞の多義性

3.1　多義の種類

本研究で扱う多義は，複合動詞の語構成要素として後項動詞が複数の意味を持つ現象を指す．複数の意味の枠組み（意味相）として，**時間相**，**空間相**そして**状態相**の3つを設定した．この意味相は，日本語の複合動詞が英語の verb particle に形式や語構成要素間の意味的制約などの観点から類似していることから，Lindner[5] が行った “out” と “up” から成る英語の verb particle の分析を参考にして決定した．

Lindner[5] は，具体物が物理的に移動する経路を表す空間的意味と，そこから拡張した抽象物の変化（時間の完了や目標の達成など）を表す意味の2つに分けて多義を説明している．本研究における意味相として，抽象物の変化の時制的意味と状態変化の意味を区別する必要性は，目的に依存すると考える．複合動詞の分析結果を自然言語処理の機械翻訳や言い換え技術に応用することを想定すると，以下の例のように各意味相における違いが明らかである．このことから，自然言語処理技術への応用に適した意味相として，**空間相**，**時間相**，**状態相**の３つの枠組みを設定することが妥当であると考える．

- 翻訳
  - 空間相：後項動詞を訳す時に particle を用いる
    例：「蹴り 上げる」→ “kick up”
  - 時間相：後項動詞を訳す時に動詞を用いる
    例：「（雨が）降り 出す」→ “begin to rain”
  - 状態相：後項動詞を訳す時副詞を用いる
    例：「あきれ 返る」→ “be thoroughly disgusted”
- 言い換え
  - 空間相：前項動詞と後項動詞を，方向とテ形を用いて言い換え可能
    例：「ボールを投げ 上げた」→「ボールを投げて 上に上げた」
  - 時間相：後項動詞を「始める」「終わる」などのアスペクト動詞に言い換え可能
    例：「酒を飲み 出した」→「酒を飲み 始めた」
  - 状態相：後項動詞を副詞「非常に」「十分に」「最後まで」などに言い換え可能
    例：「夕飯を食べ 過ぎた」→「夕飯を 必要以上に 食べた」

3.2　多義的後項動詞の選定

多義的後項動詞は，3.1 節で定義した意味相を二つ以上持ち，複数の前項動詞と接続可能な後項動詞と定義する．たとえば後項動詞「出す」について，「飛ぶ」「蹴る」などの前項動詞と結合して空間相を示し，「話す」「語る」と結合する場合は時間相の意味を持つことができる後項動詞を多義的後項動詞として扱う．本研究では慣用的表現を持つ複合動詞，たとえば「計算する」という意味の「割り出す」など，個々の語構成要素の意味が比喩的に拡張し 1 語として使われるようになった語は，辞書に複合動詞を 1 語として登録する方が効率的であると考えるため，対象から除外した．先行研究 [3][6] を参考にし，後項動詞として用いられる頻度が高く，複数の意味を持つ後項動詞を 23 語抜き出し，本研究の対象とした．

4　機械学習を用いた多義性解消実験

4.1　データの作成

機械学習を用いた多義性解消実験を行うために，毎日新聞の 1993 年度版の新聞記事 [7] をコーパスとして使用した．新聞記事 1 年分を形態素解析システム茶筌 [8] を用いて形態素に分割し，「動詞－自立」「動詞－自立」あるいは「動詞－自立」「動詞－非自立」のパターンを抽出した．抽出した複合動詞の中から形態素解析の失敗を除き，本研究の対象となる複合動詞 1094 語を抜き出した．本研究では，文脈を用いずに複合動詞の多義性を解消することを試みるために，異なり語で前項動詞 551 語，後項動詞 23 語を用いた．

前項動詞＼後項動詞	…	立てる	出す	上がる	…
⋮			⋮		
煽る			T		
焦る			T		
飛ぶ	…	?	S	S	…
叱る			?		
話す			T		
⋮			⋮		

表 1: データ形式の例

データ形式は表 1 のように，縦軸に前項動詞，横軸に後項動詞を設定し，前項動詞と後項動詞が交差する場所にコーパスから抽出した複合動詞がある場合は，3.1 節で定義した意味相を意味タグとして付与し，ない場合は記号「?」を記述した．今回は，データとして異なり語を用いたことから，前項動詞によっては 2 つの意味相を持つ場合がある．たとえば，複合動詞「打ち上げる」は「うどんを打ち上げる」は時間相に,「花火を打ち上げる」は空間相に分類される．この場合は時間相と空間相の両方の意味相を持つという「T&S(時間相&空間相)」の形式で記述した．意味タグは複合動詞の多義解消規則 [1] に基づいて人手で付与し，以下の 6 つを使用した.（意味タグ:意味相）

- T:時間相，S:空間相，A:状態相，T&S:時間相＆空間相，T&A:時間相＆状態相，S&A:空間相＆状態相

4.2 実験内容

学習には k-nearest neighbour アルゴリズムに基づいたメモリーベース機械学習システム TiMBL5.0[9] を用いた．今回実験に利用するデータ量と属性が少ないことから，少ない事例においても効率的に学習ができ，機械学習モデルを手軽に扱えるパッケージとして TiMBL を選んだ．TiMBL のパラメータは，k の値を 21，属性の重み付けを information gain，属性値の重み付けを modified value difference metic(MVDM) に設定し，10-fold による交差検定を用いて学習実験を行った．学習として与えるデータの属性を (1) 前項動詞のみ，(2) 後項動詞のみ，(3) 両動詞の意味情報，(4) 前項動詞の共起情報と後項動詞の意味情報，(5) 前項動詞の意味情報と後項動詞の共起情報，(6) 両動詞の共起情報の 6 種類に分けて実験を行い，評価結果を表 2 に示す．

対象	属性数	精度	F 値
前項動詞のみ	551	.811	.855
後項動詞のみ	23	.477	.502
両動詞（意味 × 意味）	574	.811	.856
両動詞（共起 × 意味）	574	.811	.856
両動詞（意味 × 共起）	574	.805	.850
両動詞（共起 × 共起）	574	.814	.859

表 2: 前項動詞と後項動詞を個別に属性化した評価結果

表 2 に示した通り，両動詞を属性として，前項動詞と後項動詞の属性値に共起情報を用いた場合が最も精度と F 値が高かった．F 値は，二つの意味相を持つ場合（「T&S」など）に対して，意味相を分割して個別に評価を行っているため，精度よりも高くなっている．

次に意味相別に評価したものを表 3 に示す．この結果，状態相，時間相，空間相の順に評価が高かった．空間相の精度と F 値が低かった理由として，空間相は今回対象とした多義的後項動詞における基本の意味であるが，複合動詞として用いられる時に空間的意味を持つ事例が少ないことや，空間的意味を持つかどうかは文脈に依存する部分が大きいため，共起パターンだけでは判断できないことが考えられる．今回の実験において，二つの意味相を持つのは全体の 7% の 81 個の複合動詞であるが，これらの複合動詞は共起する名詞や格の情報を用いて，個々の文脈において意味を特定する仕組みが必要となる．

クラス	精度	F 値
時間相（T）	.922	.878
空間相（S）	.845	.683
状態相（A）	.925	.924

表 3: クラス別にみた評価結果値

更に，学習に必要なデータ量と精度との関連を調べ，図 1 の学習曲線を描いた．これは，実験に使用したデータを 30 分割し，1 回の評価に 30 分の 1 ずつ加えていき，毎回 10-fold の交差検定を行った結果を示している．図 1 の曲線がほぼ平らに安定していることから，実験に使用したデータ量は本研究のタスクに対しては，充分な量のデータであることが分かった．

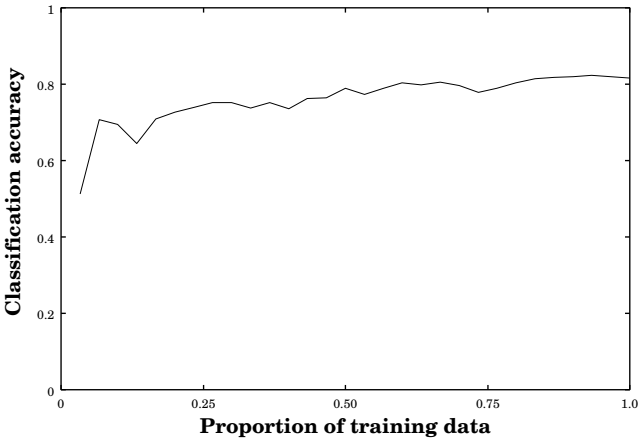

図 1: 学習曲線

以上の実験結果から，本手法は，今回利用した前項動詞と後項動詞の異なる組み合わせによる複合動詞の意味を特定するだけでなく，新規の前項動詞に対しても多義的後項動詞との共起情報がコーパスに含まれていれば，

その共起情報に基づいた意味の特定に優れていることを証明できた.

5 関連研究

本研究に関連する研究として，英語の verb particle を対象に，Levin[10] の動詞クラスを用いて，4 つの particle(down, in, out, up) と結合する動詞クラスと，生成された verb particle の意味との関連を分析した研究 [11] は，日本語の複合動詞の語構成要素間の意味的制約に共通した内容を含んでいる．また，TiMBL[9] を用いた先行研究には，英語名詞の加算名詞の分類を行った研究 [12] がある.

6 おわりに

本研究では，複合動詞を構成する後項動詞の多義性を解消するために，前項動詞と後項動詞の共起情報に基づいて学習を行った結果，約 86%の F 値となった．これにより，新規の複合動詞に対しても，コーパスから前項動詞と後項動詞の共起情報を抽出できるならば，多義的後項動詞の意味を特定できることが分かった.

今後の課題として，サポートベクトルマシン法に基づいた学習器 (TinySVM)[13] などを用いて，本研究の手法と比較を行い，より効果的な手法，属性，属性形式について検討していきたい．また，今回は意味を解析するために実験を行ったが，現在，意味タグの 1 つである「?」には，前項動詞と後項動詞が結合可能であるがコーパスに情報がない場合と，結合不可能な情報が混じっている．複合動詞の生成時の制約として，前項動詞と後項動詞との結合不可情報を加えて，生産性についても実験を行っていきたい.

参考文献

[1] Uchiyama, K. and Ishizaki, S. (2003). A Disambiguation Method for Japanese Compound Verbs. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 81-88

[2] 影山太郎 (1993). 文法と語形成 ひつじ書房.

[3] 姫野昌子 (1999). 複合動詞の構造と意味用法. ひつじ書房.

[4] 白井諭, 大山芳史, 武智しのぶ, 分部恵子, 相澤弘 (1998). “複合和語動詞に対する日英対訳用例文の収集について.” 情報処理学会第 57 回全国大会発表論文集,pp. 267-268.

[5] Lindner, S. (1981). *A lexico semantic analysis of English verb-particle constructions with UP and OUT*. Unpublished Ph.D. Dissertation, UCSD.

[6] 野村雅昭, 石井正彦 (1987). 複合動詞資料集. 国立国語研究所.

[7] 毎日新聞社 (1993). 毎日新聞 CD-ROM 版 1993 年. 毎日新聞社.

[8] 松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，湯原正幸 (2000). 日本語形態素解析システム茶筅 (Version2.2.1) 使用説明書.

[9] Daelemans, W., Zavrel, J., Sloot, K. and Van Den Bosch, A. (2003). TiMBL: *Tilburg Memory Based Learner, version 5.0 Reference Guide*. ILK Technical Report - ILK 0301, Computational Linguistics, Tilburg University.

[10] Levin, B. (1993). *English Verb Classes and Alternations - A Preliminary Investigation*. The University of Chicago Press.

[11] Villavicencio, A. (2003). Verb-Particle Constructions and Lexical Resources. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 57-64.

[12] Baldwin, T. and Bond, F. (2003). Learning the Countability of English Nouns from Corpus Data. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 463-70.

[13] Kudoh, T. (2000). TinySVM: Support Vector Machines. `http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/index.html`