語彙的複合動詞の類義表現抽出と多義別分類

神崎享子　斉藤研太　井佐原　均
豊橋技術科学大学

1. はじめに

複合動詞の複雑な意味を実際の運用の中で明らかにすることは、言語学的にも日本語教育的な観点からも重要な課題とされており、また言語処理分野においてもある表現に対する言い換え表現研究は推論やQAなどの課題の一つと考えられる。本研究では語彙的複合動詞と類似度の高い句を含めた動詞表現をコーパスから抽出し、類義表現のクラスタリングを行って、語彙的複合動詞ごとの多義別分類を行う。

語彙的複合動詞は形態的緊密性の高い複合動詞であることが知られている（影山 1993）。語としての性格の強い語彙的複合動詞は2語の構成動詞の組み合わせであるが、意味が融合し複雑な意味をもつ。2語の構成動詞がどのようなルールで意味を実現しているかについて分析した研究（影山 1993, 斉藤・石井 1997, 姫野 1999, 松本 1998, 由本 2005 など）や日本語教育の観点からの研究（森田 1978, 谷内・小森 2009, 玉岡・初 2013）、定量的な観点からの研究（石川 2010、山口 2013 など）など様々な観点から分析が行われている。一方、実際のコーパスから複合動詞の類義表現を捉え実際の文脈から実証的に意味を捉える研究は多くない。本研究は複合動詞の類義表現を抽出し多義別に分類することで、言語処理分野では言い換え表現の獲得、言語学や日本語教育分野では語彙的複合動詞の意味や動詞表現との関係を捉える一助になることを目的とする。

2. 関連研究

自然言語処理分野で日本語の複合動詞を対象にしたあいまい性解消については、Uchiyama and Ishizaki (2003), Uchiyama and Baldwin (2004) がある。あいまい性を解消し複合動詞の生成ルールを得るものであるが対象としている複合動詞は少ない。また複合的な述部表現と同義の表現を判別する泉ら（2013）の研究もある。この研究では複数の言語情報を用いており辞書の定義文なども利用している。本研究は、複合動詞2700 語を対象に行い類義表現抽出の際に国語辞典などの情報を用いない。

本研究で得る複合動詞の類義表現は多義別に分類する。今回、本研究では扱わなかったが言い換え表現をクラスタリングによって多義識別する研究として Cocos and Callison-Burch(2016)などがある。

3. データ

コーパスは 5 億文の web コーパスを使用した（Kawahara and Kurohashi 2006）。複合動詞は、国立国語研究所が公開している『複合動詞レキシコン』(2013) に収録されている約 2700 語を対象にする。評価についてはコーパスに高頻度で出現する 40 語の複合動詞を対象にして実験を行い、『複合動詞レキシコン』に収録されている「意味定義」と用例を評価基準として参照する。

4. 方法

全体のプロセスは、第一ステップとして語彙的複合動詞の類義表現候補選定のため、語彙的複合動詞と動詞（句）表現との類似度計算を行う。語彙的複合動詞ごとに取得した類義表現は、類似度上位 2000 語を取ったが適当ではない表現も多数含まれる。そこで第二ステップとして、多義を含むより適当な類義表現を絞り込むためクラスタリングを行う。第三ステップでは、多義の数は複合動詞ごとに異なるので各複合動詞の語義の数

に合った類義表現を取得する必要がある。そこでクラスタリングで絞り込んだ各複合動詞の類義表現に対して主成分分析を行い、出力結果から人手によって類義表現のクラスタを判定し、各複合動詞の多義別類義表現リストを作成、評価を行う。主成分分析による多義別表現リストはコーパスに高頻度で出現する 40 語を対象にしている。

4.1. 語彙的複合動詞と動詞（句）の類似度計算

複合動詞と動詞(句)の類似度計算のために、word2vec(Mikolov 2013)を用いて単語をベクトル表現化し、コサイン類似度を用いて複合動詞と動詞(句)間の類似度計算を行う。具体的には web コーパスの形態素解析（JUMAN）、構文解析（KNP）を行う。その際、JUMAN の辞書に登録されていない複合動詞については区切られた二つの動詞を連結し、また表記があいまいである場合のひらがな化や、動詞と修飾表現を連結した動詞句の形成などを行う。次に複合動詞も含めた動詞（句）と、係り受け関係にある名詞と格のセットを取り出し、word2vec の入力データとする。学習モデルは CBOW モデルで window の幅は 5 である。格の情報を反映させるにあたって使役や受動の除外などの例外処理も行う。

word2vec で複合動詞や動詞(句)をベクトル表現化した後、コサイン類似度を用いて複合動詞と動詞表現との類似度を計算し、複合動詞ごとにコサイン類似度の高い順から上位 2000 語の類義表現を取り出し類義表現リストを作成した。

4.2. 類義表現候補の絞込み

4.2.1 方法

4.1 節で得た類義表現リストの動詞句表現には適当ではない表現も多数含まれる。そこで多義を含んだ適当な類義表現の絞込みを行うために、クラスタ数を段階的に減らし複数回クラスタリングを行う。クラスタリングは階層型クラスタリングと k-means++を用い、結果の良い方を採用する。クラスタ数は、64、10、5 と減少させていくが、各段階で、クラスタ内の類義表現を 4.1 節で求めたコサイン類似度の高い順に並べコサイン類似度上位から 10 語を取り、それらをまとめてリスト化し、次回のクラスタリングの入力データとする。具体的には n 個のクラスタから得た上位 10 語をまとめて n*10 語のリストにして、次のクラスタリングへ入力として渡すことを繰り返す。64 クラスタでは、1 クラスタにおおよそ 30 語の類義表現を想定し、そこから上位 10 語を取ることで 64*10 語、合計 640 語のリストができる。次にこのリストを 10 クラスタで分類する。同様に各クラスタから上位 10 語を取ることで、10*10 語で合計 100 語のリストが得られる。最終的には 100 語が 5 クラスタに分類され、1 クラスタ 20 語ずつの想定になり、各クラスタの 20 語をコサイン類似度の高い順に並べ、類義表現のクラスタリング結果を得る。

4.2.2 階層型クラスタリングと k-means++との結果比較

階層型クラスタリングと非階層型の k-means++を用いて結果を比較し、本タスクに適した手法を用いる。階層型クラスタリングはウォード法を用いている。k-means++は k-means の初期値設定を工夫した手法である（Arthur 2007, 小野田他 2011）。本実験では k-means++の学習回数は 1 回である。

『複合動詞レキシコン』の意味項目を元に、ランダムに選んだ 5 語の語彙的複合動詞に対して、『複合動詞レキシコン』に登録されている「意味定義」を表す類義表現か、新規に得られた表現か、不適当な表現であるかを判定し、また、各クラスタが複合動詞の語義としてまとまっているかについても検討した。k-means++は、学習回数 1 回であったが階層型ラスタリングに比べて実行時間が長かった。初期値依存の問題を考えると複数回実行することが望ましいとされているが実行時間を考えると、階層型クラスタリングが本実験には

適当であると考えた。また、得られた類義表現は、互いに一方の手法では取れていない表現が抽出され一長一短であったが、階層型クラスタリングの方が比較的適当な表現が多く見られまとまりもよかった。以上のことから本研究では階層型クラスタリングを採用する。

4.3. 多義別の類義表現分類

4.2 節で各複合動詞に対する類義表現を絞り込み、クラスタを語義とみなして類義表現を 5 つのクラスタに分類、意味としてのまとまりについても検討した。しかし、語義の数は複合動詞によって異なる。本実験でいえばすべての複合動詞が 5 つの語義とは限らない。そこで、絞り込んだ類義表現を各複合動詞の語義の数に合ったクラスタに分類する必要がある。そこで、複合動詞ごとに、クラスタの数を事前に決定しない主成分分析をかける。入力データは、10 クラスタそれぞれからコサイン類似度の高い上位 10 語を取った合計 100 語の類義表現を対象にする。5 クラスタごとに上位 10 語を取った 50 語を対象にすると、適当と思われる類義表現が落ちてしまうため、類義表現のバラエティーがある 100 語を対象にする。

主成分分析による散布図から人手で類義表現のクラスタを判定し、複合動詞ごとにクラスタ別に分類された類義表現リストを作成した。本研究では、類義表現のクラスタは複合動詞の「語義」に相当する。今回はコーパス中に高頻度で出現した 40 語の複合動詞に対して多義別類義表現リストを作成し評価を行う。

5. 評価

対象とする 40 語の複合動詞の類義表現リストについて、自然言語処理分野の学生 4 名で被験者実験を行う。4 名を 2 名ずつ 2 グループに分け(グループ A、グループ B とする)、20 語ずつ複合動詞の類義表現リストを評価する。評価は、類義表現について『複合動詞レキシコン』の「意味定義」と用例を参照して適不適を判断する(評価 1)。次に各クラスタについて、語義が想定されるように適切に類義表現がまとまっているか(評価 2)を評価する。評価 1 の評価基準は以下の 4 点である。

(1) 登録されている意味に類義表現が適当
(2) 登録されている意味にはないが、複合動詞の類義表現として適当
(3) 登録されている意味と反対の意味
(4) 不適当な表現

同様の評価をクラスタに対しても行う(評価 2)。

6. 評価結果

グループ A、B に評価を依頼した複合動詞は 20 語ずつで、類義表現はグループ A が 527 語、グループ B が 427 語である。またクラスタ数はグループ A が 77 クラスタ、グループ B が 76 クラスタである。グループ A, B を合わせた複合動詞 40 語の類義表現の評価(評価 1)については表 1 に、クラスタの評価(評価 2)については表 2 に示す。

表 1　類義表現に対する評価　(単位　%)

類義表現として適当		反義	不適当
「意味定義」に適合	新規		
50	9	11	30

表 2　クラスタに対する評価　(単位　%)

意味として適当		反義	不適当
「意味定義」に適合	新規		
55	10	9	26

表 1 より、複合動詞の類義表現として適当と判定された表現は 59%、表 2 より複合動詞のクラスタとして適当と判定されたのは 65%となった。反義をどう捕らえるかは問題だが不適当以外の関連のある表現としては評価 1 では 70%、評価 2 では 74%となる。たとえば「持ち込む」を例にすると、複合動詞レキシコンでは「持って入る」の意味定義だけだが、提案手法では「有利に運ぶ、優位に進める」などの類義表現を取得し、有利な状

態にするという、レキシコンには記載されてない意味を想定することができる。

本研究で対象にした複合動詞40語に対して『複合動詞レキシコン』には合計64個の「意味定義」が存在する。提案手法で取得できなかった「意味定義」はそのうち14個で全体の22%存在した。「受け取る」を例にするとレキシコンの「意味定義」の「(自分宛にきたものを) 受けて、持つ」に相当する「届く、受領する、収受する」などの類義表現は取得したが、その他に収録されている「自分なりに解釈する」に相当する類義表現は取得できなかった。

7. まとめと今後の課題

本研究では複合動詞の中で特に形態的緊密性の高い語彙的複合動詞に着目して、句も含めた類義的な動詞表現の取得を行った。語彙的複合動詞は複雑な意味をもつが、本研究によって語や句を含む類義表現をコーパスから抽出することで、実際の文脈の中での複合動詞の意味をとらえる一助になると考える。本実験では高頻度で出現する40語を対象にし、提案手法により60〜70%ほどの妥当性で類義表現抽出と多義別分類を行うことができた。問題点としては、まずクラスタリング手法において階層型とk-means++を比較したが、k-means++の学習回数は本実験では1回である点、そして、最終的な複合動詞ごとのクラスタは主成分分析の散布図から人手で取得する点が挙げられる。前者についてはk-means++の学習回数を増やすと学習時間が掛かりすぎること、後者については複合動詞の数が増えると労力と時間がかかるということが問題になる。まだ問題点はあるものの、今回、計算機による実データからの抽出で、複合動詞の潜在的な意味合いを反映するような句の表現が得られた。今後、より適当な類義表現の抽出精度をあげ、人手での精査も合わせて複合動詞の類義表現をとらえていく。

(参考文献)

David Arthur (2007) *k-means++: The advantages of careful seeding*, In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete Algorithm, 1027-1035.

姫野昌子(1999)『複合動詞の構造と意味用法』ひつじ書房

石川慎一郎(2010)「現代日本語書き言葉均衡コーパス（BCCWJ）における複合動詞「〜出す」の量的分析」『統計数理研究所レポート 238』統計数理研究所

泉朋子, 柴田知秀, 齋藤邦子, 松尾義博, 黒橋禎夫(2013) 複数の言語的特徴を用いた日本語述部の同義判定. 自然言語処理 Vol.20 No.4、539-561.

影山太郎(1993)『文法と語形成』ひつじ書房

Daisuke Kawahara and Sadao Kurohashi (2006) *A Fully-lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis*, In Proceedings of HLT-NAACL2006, 176-183.

国立国語研究所『複合動詞レキシコン』(2013) http://vvlexicon.ninjal.ac.jp/

松本曜(1998)「日本語の語彙的複合動詞における動詞の組み合わせ」『言語研究』114, pp.37-77. 日本言語学会

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. (2013) *Distributed representations of words and phrases and their compositionality.* In proceedings of 27th Annual Conference on Neural Information Processing Systems, 3111–3119.

森田良行 (1978) 「日本語の複合動詞について」. 『講座日本語教育』14, 69-86.

小野田 崇, 坂井 美帆, 山田 誠二(2011) k-means 法の様々な初期値設定によるクラスタリング結果の実験的比較, 2011 年度人工知能学会全国大会（第 25 回）, 1-4.

斎藤倫明・石井正彦編(1997)『語構成』ひつじ書房

玉岡賀津雄・初 相娟 (2013)「中国人日本語学習者の語彙的複合動詞の習得に影響する要因」. 影山太郎（編)『複合動詞研究の最先端－謎の解明に向けて－』pp.413-430, ひつじ書房

谷内美智子・小森和子 (2009) 「第二言語の未知語の意味推測における文脈の効果－語彙的複合動詞を対象に－」. 『日本語教育』 142, 113-122.

Kiyoko Uchiyama and Shun Ishizaki. (2003) *The Method on the Semantic Analysis for disambiguation of compound verbs*. In proceedings of the 9th annual conference of Natural Language Processing,163-166.

Kiyoko Uchiyama, Timothy Baldwin,.(2004) *Automatic Disambiguation of Compound Verbs by Machine Learning*. In proceedings of the 10th annual conference of Natural Language Processing, 741-744.

山口昌也(2013)「複合動詞「〜込む」と前項動詞の格関係」影山太郎編『複合動詞研究の最先端』ひつじ書房

由本陽子(2005)『複合動詞・派生動詞の意味と統語』ひつじ書房