# コーパスから算出した語の親和性によって構文パターンの曖昧性を解消する試み

野澤 元[1]　河原 大輔[2]

1. 京都外国語大学 2. 京都大学

E-mail: h_nozawa@kufs.ac.jp, dk@i.kyoto-u.ac.jp

**あらまし**　英語の特定の構文の解析精度を上げるための手がかりとして、別の構文における語の分布パターンを語の親和性として算出し、利用できないかについて調査した。

**キーワード**　格フレーム辞書、構文解析、語彙のサブカテゴリ化

## 1. はじめに

本研究では、英語の特定の構文の解析精度を上げるための手がかりとして、別の構文における語の分布パターンを語の親和性として算出し、利用できないかについて調査した。具体的には、N1 V1 N2 to V2 という構文の事例における、V1 に対する V2 の to 不定詞の統語的関係が、1.目的格補語である場合と、2.目的を表す副詞句である場合を区別するために、コーパス内において V2 の動詞が N2 の名詞を主語とする頻度を手がかりとして利用できる可能性を調べた。

## 2. 格フレーム辞書での解析誤り

格フレーム辞書は、動詞にどのような格が、どのような語を伴って出現するのか記述した辞書である。格フレーム辞書は、FrameNet (Baker et al., 1998)のように人手で構築されたものと、大規模コーパスから自動的に構築されたもの(Kawahara et al., 2014)がある。自動構築された格フレーム辞書は、大量の用例を参照できるため、さまざまな言語処理アプリケーションでの利用が期待できる。しかし、このような辞書はタグ付けされていない大規模コーパス(the English Gigaword corpus)を用いているため、その結果には様々な誤りが含まれる。その中の一つが、語の表層分布だけでは判別のできない、曖昧な構文の解析結果である。例えば、(1)と(2)は、N1 V1 N2 to V2 という構文パターンを共有しているが、(1)における to eat that fruit は文の主動詞である entice の目的格補語であるのに対し、(2)における to eat the fruit.は目的を表す副詞句である。しかし、現在の格フレーム辞書では、(2)の不定詞は(1)のものと同様に、目的格補語として誤って解析されている。

1. Satan enticed Eve to eat that fruit.
2. Most people buy the tree to eat the fruit.

本研究では、この N1 V1 N2 to V2 という構文パターンにおける V2 の to 不定詞の、目的格補語と目的を表す副詞句という二つの用法を分離するために、その手がかりとしてコーパス内の N2 と V2 の分布パターンが利用できないかを調査した。

## 3. to 不定詞の二つの用法

ある動詞が直接目的語と目的格補語の to 不定詞をとる場合、直接目的語は to 不定詞の「意味上の主語」と呼ばれるように、しばしば to 不定詞の動詞が表す行為の動作主を表す。そのため、例えば、(1)の N2 である名詞 Eve と、V2 である動詞 eat をそれぞれ主語と述語とする(3)のような文は、よく見られることが予想される。

3. Eve eats * fruit(s).
4. * tree(s) eat(s) * fruit(s).

これに対して、(2)の N2 ある名詞 tree と、V2 である動詞 eat をそれぞれの主語と述語とする(4)のような文は、比較的稀だと考えられる。したがって、このようにコーパス内において V2 の動詞が N2 の名詞を主語とする頻度を調べることによって、もと文における V2 の to 不定詞が目的格補語なのか、それとも目的を表す副詞句なのかを、より正確に判別することが期待される。以下、この頻度を N2 と V2 の親和性を表す指標と見なし、「N2-V2 親和性」と呼ぶ。

## 4. to 不定詞の用法の判別

N1 V1 N2 to V2 という構文パターンにおける、to 不定詞の二つの用法を区別するには、大きく分けて二つの段階が考えられる。まず第一段階では、V1 の種類によって to 不定詞の用法を判別する。もし、V1 がそもそも目的格補語の to 不定詞をとることのできない動詞であれば、V2 の to 不定詞は自動的に目的を表す副詞句だと判定できる。例えば(2)の場合、V1 である動詞 buy は目的格補語の to 不定詞をとることが出来ないため、V2 の to 不定詞は目的を表す副詞句だと判定できる。ただし、このような判別の基準として、どの動

詞は目的格補語の to 不定詞をとることができるのかを教える辞書が必要となる。

第二段階では、目的格補語の to 不定詞をとることのできる動詞を V1 とする事例について、それぞれの V2 の to 不定詞がどちらの用法であるのかを判別する。例えば、(5)と(6)は N1 bring(V1) N2 to V2 という構文パターンを共有しているが、V2 の to 不定詞の用法は異なっている。

5. Sadly this media focus brings people to believe that dieting is necessary.
6. She had also brought the recipe to cook for us tonight.

(5)における V2 の to 不定詞は動詞 bring の目的格補語であるのに対し、(6)では目的を表す副詞句である。これらの二つの用法は、すでに述べたように、N2-V2 親和性を手がかりとして区別できると期待される。

## 5. 動詞のサブカテゴリ化

第一段階で必要とされる辞書は、従来では語彙のサブカテゴリ化のタスクとして構築されてきた。例えば、Korhonen et al. (2006)は統語解析されたコーパス中の文を自動的に 163 種類のサブカテゴリ化フレームに分類している。ただし、本研究の課題には、それほど多様なフレームの情報は必要ないこと、また、一つの手法の有用性を広く検証するという目的から、N2-V2 親和性を利用することで、to 不定詞の目的格補語の用法に限定したサブカテゴリ化辞書を構築できる可能性を調べた。

まず、約 300 万文を含む大規模コーパス(the English Gigaword corpus の一部)の中から、N1 V1 N2 to V2 の構文パターンをとる文を抽出した。次に、それぞれの事例における N2-V2 親和性を計算した。この際に、N2 が人称代名詞であるものと、固有の人名であるものは除いた。これは N2 と V2 がそれぞれ、単に人物であることのみを表す名詞(または代名詞)と、典型的に人間の行為を表す動詞の組み合わせである場合、V2 の動詞が N2 を主語としてとる確率が非常に高くなり、N2-V2 親和性をあまりにも高く見積もってしまうためである。これらの事例を取り除いた結果、最終的には分析対象として 19242 文の事例が得られた。そしてさらに、V1 のそれぞれの動詞毎に、N2-V2 親和性の平均値を求めた。その結果が、表 1 である。

ここには、事例数が 10 以下の動詞は除かれている。実際にそれぞれの動詞が、目的格補語の to 不定詞をとるかどうかは、プログレッシブ英和・和英中辞典を参照して人手で確認し、1 は用法あり、0 は用法なし、0.5 は to 不定詞が be 動詞の場合のみ用法ありという形で

| V1 | N2-V2 親和性 | V2 目的格補語 |
|---|---|---|
| understand | 0.371 | 1 |
| like | 0.320 | 1 |
| lose | 0.318 | 0 |
| want | 0.310 | 1 |
| pressure | 0.293 | 1 |
| become | 0.288 | 0 |
| check | 0.277 | 0 |
| expect | 0.277 | 1 |
| put | 0.263 | 0 |
| convince | 0.256 | 1 |
| get | 0.249 | 1 |
| encourage | 0.248 | 1 |
| undertake | 0.248 | 0 |
| warn | 0.245 | 1 |
| represent | 0.236 | 0.5 |
| urge@PAS | 0.222 | 1 |
| hire | 0.220 | 0 |
| lead@PAS | 0.213 | 1 |
| add | 0.212 | 0 |
| allow | 0.210 | 1 |
| hope | 0.209 | 0 |
| invite | 0.202 | 1 |
| enable | 0.200 | 1 |
| force | 0.199 | 1 |
| promise | 0.198 | 0 |
| ask | 0.192 | 1 |
| lead | 0.184 | 1 |
| recognize | 0.184 | 0.5 |
| persuade | 0.179 | 1 |
| tell | 0.179 | 1 |
| carry | 0.172 | 0 |
| help | 0.165 | 1 |
| oblige | 0.165 | 1 |
| advise | 0.163 | 1 |
| reserve | 0.163 | 0 |
| leave | 0.162 | 1 |
| urge | 0.161 | 1 |
| permit | 0.159 | 1 |
| request | 0.157 | 1 |
| cause | 0.152 | 1 |
| protest | 0.151 | 0 |
| demand | 0.150 | 0 |
| choose | 0.149 | 0.5 |
| compel | 0.147 | 1 |
| press | 0.143 | 1 |
| instruct | 0.141 | 1 |
| return | 0.141 | 0 |
| allow@PAS | 0.138 | 1 |
| know | 0.132 | 1 |
| secure | 0.130 | 0 |
| do | 0.127 | 0 |
| be | 0.125 | 0 |
| submit | 0.124 | 0 |

表 1

符号化した。

結果を見ると、必ずしも厳密な対応関係はないものの、大まかな傾向としては、N1 V1 N2 to V2 という構文パターンの V1 として現れる動詞は、N2-V2 親和性の平均が高いほど、目的格補語をとる可能性が高いことが示された。

## 6. 曖昧な構文の判別

第二段階では、目的格補語の to 不定詞をとることのできる動詞を V1 とする事例について、それぞれの V2 の to 不定詞が、V1 の目的格補語であるのか、それとも目的を表す副詞句であるのかを判別する。すでに述べたように、ここでは N2-V2 親和性を判別の手がかりとして用いる。ただし、直接目的語と目的格補語の to 不定詞をとる構文が、その主要な用法の一つであるような動詞については、N1 V1 N2 to V2 の構文パターンで用いられた場合、そのほとんどにおいて、V2 の to 不定詞は目的格補語の用法である。表 1 に見られるように、そのような動詞は平均の N2-V2 親和性が高いだけでなく、他と比べて事例数も多い。例えば、全ての事例に占める割合は、allow が 9.22%、want が 3.1%、expect が 2.73%である。

したがって、現実的に第二段階の判別が必要なものは、直接目的語と目的格補語の to 不定詞をとる構文が、その主要な用法ではない動詞の事例である。まだ全ての事例の確認が済んでいないが、例えば、表 2 に示すように動詞 bring の場合は、N2-V2 親和性が高いほど V2 の to 不定詞が目的格補語であるという傾向が見られる。

| N1 | V1 | N2 | V2 | N2-V2 親和性 | V2 目的格補語 |
|---|---|---|---|---|---|
| | bring | government | respect | 0.097 | 1 |
| seven | bring | number | defect | 0.046 | 0 |
| | bring | state | exchange | 0.040 | 1 |
| tie | bring | voter | support | 0.021 | 1 |
| | bring | party | pull | 0.015 | 1 |
| | bring | document | justify | 0.007 | 0 |
| | bring | pressure | change | 0.003 | 0 |
| | bring | pressure | end | 0.003 | 0 |
| group | bring | tourist | satisfy | 0.003 | 0 |
| | bring | korea | agree | 0.002 | 1 |
| | bring | court | face | 0.001 | 0 |
| | bring | cigarette | market | 0 | 0 |
| | bring | protagonist | see | 0 | 1 |
| dance | bring | lever | win | 0 | 0 |
| guard | bring | scrutiny | make | 0 | 0 |
| he | bring | popularity | play | 0 | 0 |
| investment | bring | skill | boost | 0 | 0 |
| meeting | bring | historian | formulate | 0 | 0 |
| rich | bring | carpet | auction | 0 | 0 |

表 2

## 7. 議論

第一段階での動詞のサブカテゴリ化では、目的格補語をとることができないにもかかわらず、N2-V2 親和性が高い動詞がいくつか見られた。それらの事例を調べると、これにはいくつかの原因が見られた。

そのうちの一つは、特定の動詞を V1 とする事例の数が少なく、またそれらがたまたま大きな割合で、N2-V2 親和性が高くなるような根本的に誤った構文解析の結果を含んでいる場合である。動詞 lose はこれに該当し、事例数は 30 であるが、そのうちの 5 つの事例は誤った解析の結果、N2-V2 親和性が非常に高かった。この問題については、より大きなサイズのコーパスを用いることで改善が見込まれる。

より興味深いのは、動詞の意味的な性質が原因となっている場合である。N1 V1 N2 to V2 の構文パターンにおいて、V1 が動詞 hire である事例の多くでは、(7) のように、V2 の to 不定詞は目的を表す副詞句である。

7. Miller has had to hire a promotions firm to manage endorsements.

しかし、この場合の文の意味は、基本的には、本来は N1 が V2 をしたいと思っているが、何らかの理由で出来ないため、N2 に金を払って代行してもらうということである。そのため、この文は N2 が V2 を行うこと

ができるということを、その前提とするのである。その結果、このような事例では V2 の to 不定詞は目的を表す副詞句であるにもかかわらず、全般的に N2-V2 親和性が高くなるのである。この問題は、現在の手法だけでは解決することができない。

第二段階での V2 の to 不定詞の曖昧性の除去では、V1 の動詞によっては、ある程度の傾向が見られたものの、必ずしも明確な結果は得られなかった。曖昧性の除去が必要な事例は、現在のサイズのコーパスでは十分な数がが得られないため、今後より大きなサイズのコーパスを用いて、再度検討したい。

## 8. 結論

本研究の結果は、その精度についてはかなり改善の必要があるが、N2-V2 親和性を手がかりとして、N1 V1 N2 to V2 という構文における、V1 に対する V2 の to 不定詞の統語的関係が、1.目的格補語である場合と、2.目的を表す副詞句である場合を区別するために利用できる可能性を示したと言える。今後はコーパスのサイズを拡大し、また、その他の指標を組み合せることにより、さらなる精度の向上を目指す必要がある。

## 文　献


Baker, Collin F., Charles J. Fillmore and John B. Lowe. (1998). The Berkeley FrameNet project. In *Proceedings of the COLING-ACL*, Montreal, Canada.

Kawahara, Daisuke, Daniel W. Peterson, Octavian Popescu and Martha Palmer. (2014) Inducing Example-based Semantic Frames from a Massive Amount of Verb Uses, In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (EACL2014), Gothenburg, Sweden. (to appear)

Korhonen, Anna, Yuval Krymolowski and Ted Briscoe. (2006) A Large Subcategorization Lexicon for Natural Language Processing Applications, In *Proceedings of the 5th International Conference on Language Resources and Evaluation*. (LREC2006), Genoa, Italy.