# 交替関係を利用した結合価辞書の獲得方法

藤田 早苗　　Francis Bond
{sanae, bond}@cslab.kecl.ntt.co.jp
NTT コミュニケーション科学基礎研究所

## 1 はじめに

用言の必須格や選択制限などの情報 (以降、結合価情報と呼ぶ) は、機械翻訳や要約、QA など自然言語処理のほとんどの分野で有用である。しかしこういった豊富な情報をもつ辞書は、人手で作成するとコストと時間がかかり、自動的に作成すると精度が保証できないという問題がある。

しかし、交替の情報があれば、一つのエントリから複数の結合価エントリが作成できるといわれている (Dorr, 1997)。そこで、「溶ける」対「溶く」のような自動詞対他動詞の交替を対象に、交替の片側の結合価情報からもう片側の結合価情報を獲得する実験を行なう。実験では自動詞（他動詞）から他動詞（自動詞）の結合価情報を獲得し、既存の結合価情報と比較する。また、獲得する結合価情報の一部について翻訳試験による評価を行い、交替情報から新しい語の結合価情報が獲得ができるかの検証を行なう。

実験には、結合価情報を持つ辞書として、NTT で日英機械翻訳システム**ALT-J/E** 用に開発してきたパターン対辞書 (池原他, 1997) を利用する。

## 2 日本語と英語の交替

日本語と英語の交替を比較するため、Jacobsen (1981) の交替リストを元にして、日本語の自動詞対他動詞の交替とその英訳のリストを作成した。このリストには、日本語と英語の見出し語情報が含まれる。例えば [**自動詞**] 「溶ける」 *dissolve* 対 [**他動詞**] 「溶く」 *dissolve* のような情報である。「溶ける」「溶く」の結合価情報は図 1のようになっている。ここで他動詞の「ガ格」を **A**、「ヲ格」を **O**、自動詞の「ガ格」を **S**とすると、**O**と**S**が対応し、**A**は自動詞側では消えている。

このリストの英訳の構造を変えるだけで同じ英訳を作成できるかどうかを調査し、交替のタイプを分類した (表 1)。同じ日本語組合せに対し、英訳組合せが複数ある場合「作成可能」な組合せが一つでもあればそちらに分類した。表 1で「作成不可能」に分類されたものは、英語が全く異なるなど、元の英訳の構造を変えるだけでは同じ英訳を作成できない。表 1 から、交替の片側のパターン対を作成する時に元の英訳から交替の英訳を作成できるのは、最大 51.0% である。但し、「作成可能」に分類されたものでも、最適な英訳が作成できるとは限らない。また、「作成不可能」に分類されたものでも、他のエントリや辞書から異なる英訳を抽出すれば「作成可能」になる可能性がある。

## 3 交替と辞書

パターン対辞書には自動学習によって交替リンクが張られているものがある (Bond 他, 2002)。そこで、本実験では人手作成の交替リストにあり、交替リンクが張られている 311 組のパターン対（字面の組合せは 144）を対象とする。

図 1の結合価情報はパターン対辞書から取り出した物である。ここで、N1,N2 等は主格や目的格を表す変数である。また、《 》で示したのは格の選択制限であり、意味属性のリストで与えられる。また、パターン対辞書の英語側は英語の字面や格要素を定義する**肉付部分**と、品詞や文型を定義する**骨格構造**が組になっている (横尾他, 1994)。**骨格構造**の種類は 616 種類だが、上位 10 種類 (***A Vt O*** や ***S be Adj*** 等) で全体の 72% 以上をカバーする。

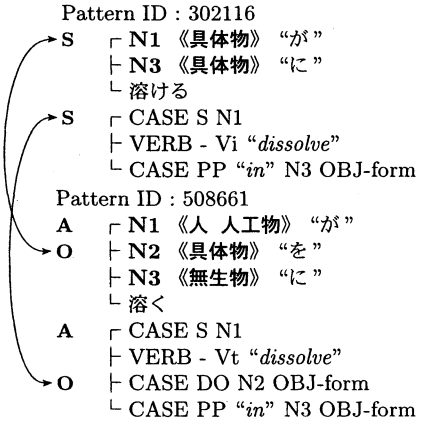


図 1: 結合価情報 (パターン対辞書) の例 溶く ⇔ 溶ける

表 1: 交替タイプ分類

| 日本語 | | 英訳 | | 英語構造 | | 数 | (%) |
|---|---|---|---|---|---|---|---|
| [自動詞] | [他動詞] | [自動詞] | [他動詞] | [自動詞] | [他動詞] | | |
| 作 | 弱まる | 弱める | S weaken | A weaken O | *S Vi* | *A Vt O* | 154 | 27.6 |
| 成 | 漏れる | 漏らす | S be omitted | A omit O | *S be Vt-ed* | *A Vt O* | 100 | 17.9 |
| 可能 | 泣く | 泣かす | S cry | A make O cry | *S Vi/be Adj* | *A Vc O Vi/Adj* | 30 | 5.4 |
| 作成 | 亡くなる | 亡くす | S pass away | A lose O | *S Vi* | *A Vt O* | 270 | 48.4 |
| 不可能 | 怒る | 怒らす | S get angry | A anger O | *S Vc Adj* | *A Vt O* | 4 | 0.7 |

Vc は *make,get,let,become*等の制御動詞。 実際のパターンには随格が含まれることもある。

## 4 パターン対エントリ作成方法

本実験では「象は鼻が長い」のような「ハ格」と「ガ格」を両方含むパターン対は対象外とする。

### 4.1 日本語側作成方法

他動詞から自動詞を作成するには、他動詞の**A**を削除し、**O**の選択制限を自動詞の**S**の選択制限にする。それ以外の格はそのまま複製する。

自動詞から他動詞を作成するには、自動詞の**S**の選択制限を他動詞の**O**の選択制限にする。更に、他動詞の**A**として「N1《**主体**》が」を追加する。意味属性を《**主体**》にするのは、既存の辞書で「N1が」との組合せで出現頻度が最も高いからである。但し、自動詞側に「ニ格」があり、かつ、英訳が *by* のものは、「ニ格」の選択制限を**A**の選択制限にし、他動詞の「ニ格」は複製しない。例えば、「N1《**主体 動物**》が N3《*》に 驚く」 *N1 be surprised at/by N3* から他動詞を作成すると、「N1《*》が N2《**主体 動物**》を 驚かす」 *N1 make N2 surprised* になる。それ以外の格はそのまま複製する。

### 4.2 英語側作成方法

まず**肉付部分**を作成し、**肉付部分**と一致する**骨格構造**を選択する (図 2)。**骨格構造**が一意に決まらない場合は、候補のうち、その交替タイプで最も使用頻度の高い**骨格構造**を利用する。それでも一意に決まらない場合は全パターン中で最も使用頻度の高い**骨格構造**を利用する。但し、一致する**骨格構造**がない場合は典型的な**骨格構造**を利用[1]する。

## 5 作成結果と議論

交替リンクが張られている 313 組のパターン対（字面の組合せは 144）を用いてパターン対を作成した。但し、1 つのパターン対に複数のパターン対からリンクが張られている事もあるため、自動詞側異なりが 237 パターン、他動詞側異なりが 265 パターンである。

[1]自動詞側では *S Vi*あるいは *S be Vt-ed/Adj*、他動詞側では *A Vt O* の構造

### 5.1 既存パターン対辞書との比較

既存の辞書の交替パターン対を $P_I$,$P_T$ とし、$P_I$ を用いて作成したパターン対を $N_T$、$P_T$ を用いて作成したパターン対を $N_I$ とすると $N_I$ と $P_I$、$N_T$ と $P_T$ を比較する。ここで、$P_I$,$N_I$ は自動詞、$P_T$,$N_T$ は他動詞である。表 2では、各項目について $N_I$ と $P_I$、$N_T$ と $P_T$ が一致する数を数えた。表 2で「日本語側すべて一致」としたものは、選択制限も含めて全く同じパターン対を作成できたものである。「英語側すべて一致」としたものは、選択制限は考慮していないので、日本語側の「N1,N2,…と格助詞が一致」に対応する。

表 2: 既存パターン対との比較 ($N_I$ 対 $P_I$,$N_T$ 対 $P_T$)

| | | [自動詞] | | [他動詞] | |
|---|---|---|---|---|---|
| | 比較項目 | 一致数 | (%) | 一致数 | (%) |
| 日 | N1,N2,… | 113 | 43.5 | 134 | 56.5 |
| 本 | 格助詞 | 88 | 33.8 | 99 | 41.8 |
| 語 | N1,…と格助詞 | 83 | 31.9 | 93 | 39.2 |
| 側 | すべて | 30 | 11.5 | 17 | 7.2 |
| 英 | N1,N2,… | 112 | 43.1 | 124 | 52.3 |
| 語 | 骨格構造[2] | 50 | 19.2 | 79 | 33.3 |
| 側 | 主辞 | 50 | 19.2 | 61 | 25.7 |
| | すべて | 29 | 11.2 | 30 | 12.7 |
| | 日英すべて | 11 | 4.2 | 5 | 2.1 |
| | 作成数 | 260 | | 237 | |

[2] 典型的な**骨格構造**を利用したものは、他動詞作成側で 17 個、自動詞作成側で 15 個あった。

表 2から、(1) 日本語側の方が英語側より既存のパターン対と同じように作成できている割合が高い、(2) 既存のパターン対と同じように作成できた割合は低い、ということが分かる。

(1) の原因は、日本語側の交替を元にパターン対を作成しているからである。2章で述べたように、日本語が交替するときに英語側も同じ英語で交替が作成できるのは最高 51% であり、英語側の方が作成できない確立は高い。

また (2) の原因には、(a) 既存の交替に元々選択制限や格助詞の差がある、(b) 既存の交替に元々 N1,N2,…の対応にばらつきがある、(c) 交替リンクが自動作成したものであるため、本来交替でないも

[**自動詞側**] 作成方法：

- 制御動詞 (*make,have,get,cause*) を取る**骨格構造**の場合
  - **A** Vc **O** Vt/Adj ⇒ **S** be Vt-ed/Adj (1 例 0.4%)
- 制御動詞を取る**骨格構造**ではない場合
  - 他のパターン対に自動詞の用法がある場合
    - **A** Vt **O** ⇒ **S** Vi (162 例 62.3%)
  - 他のパターン対に自動詞の用法がない場合
    - **A** Vt **O** ⇒ **S** be Vt-ed (94 例 36.2%)

例外として、主辞が制御動詞でない *have*の場合 *There is* 構文にする。(3 例 1.2%) 例えば、「及ぼす」 *N1 have N2 on N3* ⇒ 「及ぶ」 *There be N1 on N3*

[**他動詞側**] 作成方法：

- 元の**骨格構造**が**S** Vi の場合
  - 他のパターン対に他動詞の用法がある場合 ⇒ **A** Vt **O** (102 例 43.0%)
  - 他のパターン対に他動詞の用法がない場合 ⇒ **A** Vc **O** Vi (21 例 8.9%)
- **S** be Adj ⇒ **A** Vc **O** Adj (40 例 16.9%)
- **S** be Vt-ed ⇒ **A** Vt **O** (43 例 18.1%)
- **S** Vt ⇒ **A** Vt **O** (31 例 13.1%)

但し、 Vc は *make*を利用

図 2: 英語作成方法

のも含まれている、等が上げられる。

原因 (a)(b) については、既存の交替の差に意味があるかどうか、という問題がある。例えば図 1 に示した交替でも、「ニ格」の選択制限が「溶く」では《**無生物**》、「溶ける」では《**具体物**》と違いがある。ここで、図 1の《**具体物**》は《**無生物**》の一つ上のノードであり、子供には《**生物**》もあるため、選択制限は「溶ける」でも《**無生物**》の方が適切である。つまり、原因 (a)(b) については全く同じように作成するより、意味のある差なのかを検討し、より適切な結合価情報へ直す事が重要である。このように、既存の辞書との比較だけでは、よいパターン対を作成できたかどうかの判断が難しいため、翻訳により評価を行なった (6章)。

### 5.2 A O S の意味属性の差分の調査

交替であっても自動詞と他動詞では選択制限は微妙に異なり、その差は意味があると思われる。そこで、 Bond 他 (2002) が学習で作成した自動詞対他動詞の 449 組の交替を使って自動詞の**S**、他動詞の**A**、**O**の意味属性の日本語語彙大系におけるレベルを数えた (図 3)。レベルは上位になるほど意味が一般的で、下位になるほど特殊化する[3]。図 3によると、**A** は《**主体**》や《**具体物**》などが含まれるレベル 2 が突出して多く、**S**と**O**は傾向は似ているが、**O**の方がレベルが低い物が多い。

また、**S**、**A**、**O**の意味属性のうち、《**主体**》配下の意味属性は**A**と**S**が 17%、**O**が 8% だった。つまり、**S**の方が**O**よりレベルが若干高いものが多く、《**主体**》や《**主体**》配下が多い。今回はそういったことを考慮せず、選択制限をそのまま複製したので、他動詞から自動詞を作成した場合には、**S**の選択制限が狭くなり、自動詞から他動詞を作成した場合には、**O**の選択制限が広くなる傾向がある。

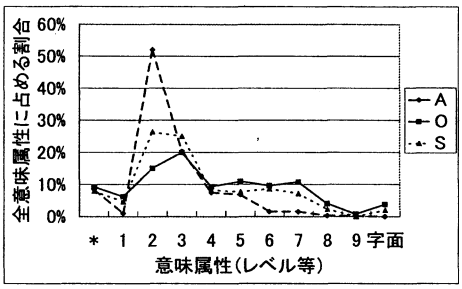


図 3: 意味属性のレベル

## 6 翻訳による評価と議論

### 6.1 評価

本実験で対象とした交替の見出し語が既存のパターン対辞書にいくつ登録されているかを調査すると、1 パターンだけ登録されていたものが 25 見出し語あり、1–7 パターン登録されているもので、交替の見出し語の半数以上をカバーした。

見出し語が既存のパターン対辞書にない場合に、交替情報から獲得する場合を考え、既存のパターン対辞書に 1 パターンだけ登録されている 25 見出し語を対象に翻訳結果を比較評価した。評価対象は、自動詞側で「亡くなる」「混ざる」等 10 語、他動詞側で「亡くす」「閉める」等 12 語である。

自動詞側結果を評価する場合は、翻訳は、元のパターン対辞書を利用した場合 ($P_I$ を含む)、作成したパターン対 ($N_I$) を含む辞書を利用した場合、パターン対辞書から $P_I$ を削除した場合の 3 通り行なった。試験文には例文と新聞[4]からその見出し語を使っている文を抽出し、翻訳結果に $N_I$ か $P_I$ が使わ

[3]図 3の「*」は何でも取り得るという意味。「字面」は選択制限を字面で与えている。

[4]毎日新聞'91-2000 と日本経済新聞'90-99 を利用。

れている文だけを1見出し語につき最大5文残し[5]、評価文とした。他動詞側でも同様に行なった[6]。

$P_I$ や $P_T$ を削除した場合と元のパターン対辞書を利用した場合の比較結果と、$P_I$ や $P_T$ を削除した場合と $N_I$ や $N_T$ を利用した場合の比較結果を図4に示す。

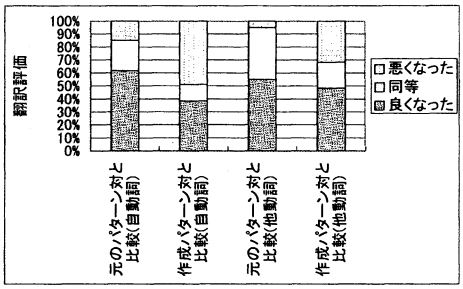


図 4: 翻訳評価結果 (パターン対を削除した場合との比較)

### 6.2 議論

図4によると、元のパターン対の方が作成したパターン対を使うより翻訳精度は高い。しかし、自動詞側と他動詞側の結果を平均すれば、パターン対がないよりは、作成したものがある方がよくなる。パターン対がない場合に比べて交替から作成したパターン対を使う場合は自動詞側で38.3%、他動詞側で48.3%翻訳結果が良くなっている。更にこれに「同等」も加えると自動詞側で51.0%、他動詞側で68.3%を占める。2章より、元の英訳を利用して規則的に正しい英訳を作成できる可能性は最大51%なので、良く作成できている。

作成したパターン対による翻訳の方が悪くなった場合の原因は、(1) 元の英語の品詞 (過去分子形か、形容詞か) が間違っている、(2) 自動詞や他動詞としての用法があるかどうかの判断が間違っている、(3) 「亡くす」 *lose* 対「亡くなる」 *pass away* のように、同じ英語では適切な交替を作成できない、等があげられる。

原因 (1) によるものは、元のパターン対の品詞情報を正しくすれば、人手で作成したパターン対と比べても、評価がそう劣らないものができる。逆に、英語主辞の字面は同じなのに交替パターンがうまく作成できない場合、辞書の品詞が (過去分子形か、形容詞か) 間違っている場合が多い。品詞の判断は非常に難しい場合があるが、問題発見の手がかりとなる。

[5]5文に満たなかったのは、「被さる」の2文である

[6]自動詞側で47文、他動詞側で60文試験文がある。

原因 (2) の例をあげる。[**自動詞**] 「仕上がる」 *N1 be finished* 対 [**他動詞**] 「仕上げる」 *N1 finish N2* の交替では、自動詞側で作成した英訳は *N1 finish* となった。これは、他のパターン対で *N1 finish in N3* という自動詞の用法があるためである。しかし *finish* は単独では自動詞としては用いられないため、英訳が悪くなっている。

また、他動詞側で制御動詞 *make* を用いて英訳を作成したのは61パターン (25.7%) で、表1の5.4%の5倍近い。これには、(1) 本来自動詞の用法があるのに、他のパターン対に自動詞の用法が見つからず制御動詞を用いている、(2) 英語側は違う語を用いて交替を作成すべきだが元の英語を用いて強引に英訳を作成した、等の理由があげられる。

このように、英語側の作成精度をより高くするには、特別な前置詞と一緒に使うときだけの用法や、語義による用法の異なり、英語で交替できる語どうか等、より詳しい英語側の情報が必要である。

## 7 まとめ

自動詞対他動詞の交替を用いて交替の片側のパターン対からもう片側のパターン対を作成し、既存のパターン対と比較した。また、対象見出し語が元のパターン対辞書に1つしか登録されていない場合に翻訳評価を行なった。その結果、翻訳結果が良くなったものと同等のものを加えると、英語側交替を作成できる予測値以上に良いパターン対を作成できた。今後は、他の言語資源から英語側の交替情報をより豊富にし、(1) 既存の辞書にない語のパターン対の獲得、(2) 他の交替の調査、(3) 日本語と英語の交替の差の調査、等を行なう。

## 参考文献

池原 悟, 宮崎 雅弘, 白井 諭, 横尾 昭男, 中岩 浩巳, 小倉 健太郎, 大山 芳史, 林 良彦. 日本語語意大系. 岩波書店, 1997.

横尾 昭男, 中岩 浩巳, 白井 諭, 池原 悟. 日英機会翻訳用スケルトン - フレッシュ型構文意味辞書の構成. 情報処理学会第 *48* 回全国大会, pages 139–140, 1994.

Francis Bond, Timothy Baldwin, 藤田 早苗. Detecting alternation instances in a valency dictionary. 言語処理学会第 *8* 回年次大会, pages 519–522, 2002.

Bonnie J. Dorr. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12(4):271–322, 1997.

Wesley Jacobsen. *Transitivity in the Japanese Verbal System.* PhD thesis, University of Chicago, 1981. (Reproduced by the Indiana Univeristy Linguistics Club, 1982).