# 対訳コーパスを用いた同義述語獲得の検討

泉　朋子† Wong Ben Kin Sum‡ 齋藤　邦子† 松尾　義博†
†日本電信電話株式会社　NTTメディアインテリジェンス研究所
{izumi.tomoko, saito.kniko, matsuo.yoshihiro}@lab.ntt.co.jp
‡ The University of British Columbia　Faculty of Applied Science
ben.wong@alumni.ubc.ca

## 1　はじめに

現在、Web上のブログや音声対話ログなど大量のテキストから欲しい情報を探し出す検索技術や、有益な情報のみを自動で抽出・集計するテキストマイニング技術の高精度化が求められている。これらを実現するために必要なのが、計算機による自然文の意味理解である。

例えば、下記のような2つの文があった場合、これらが「同じ事を表している」と判別できなくては、利用者が求める情報を正しく検索できなかったり、テキストマイニングで必要な「同じ情報のまとめ上げ」ができない。

(1)　I bought a smartphone.
(2)　I purchased a smartphone.

特に、”bought”や”purchased”といった述語は、文の核情報を表しており、これらが同義であるという識別が自動で可能になれば、より精度の高い情報検索・情報抽出が可能となる。

これらを可能にするためのひとつの方法として、同義語辞書の自動獲得がある。例えば、”purchase”の同義語として”buy”が入っている辞書があれば、その辞書を用いて、「purchase と buy は同義である」と識別することが可能である。

本稿では、対訳コーパスを用いて、述語の同義語辞書（同義述語辞書）の自動獲得を検討する。(1)と(2)は、日本語では「私は、スマートフォンを買った。」というように、同じ表現(i.e.,「買う」)に訳すことが可能である。この「他の言語で同一の単語に訳されている表現は同義と成りうる」という特徴に着目し、対訳コーパスからの同義述語辞書の獲得を行う。さらに、対訳コーパスの情報に加えて、述語に係る項を特徴とした類似度を組み合わせることで、同義述語獲得の精度を向上させる。なお、本稿では英語を対象とする。次節では、既存研究について述べる。3節では提案手法について説明し、4節で実験、5節で考察を行う。

## 2　既存研究

単一言語のパラレルコーパスを用いた同義表現の獲得として、Barzilay & McKeown (2001)がある。彼らは、同一の小説に対し、複数の翻訳者によって翻訳されたデータを用いて、言い換え表現の抽出を行っている。例えば、下記のような例である。

(3) Emma **burst into tears** and he tried to **comfort** her, saying things to make her smile.
(4) Emma **cried**, and he tried to **console** her, adorning his words with puns.

(3)と(4)は、もともと同じ事を表す一文を訳した文である。そのため、これら2文間の文字列に違いがある場合、それらは「同義表現」の可能性が高い。Barzilay & McKeown(2001)は、この特徴を用いて”burst into tears”と”cried”、”console”と”confort”のような同義表現の獲得を行った。同様に、Shinyama et al. (2002)は、同じ出来事に対して書かれた複数の新聞記事を用いて、言い換え表現の獲得を行っている。しかし、単一言語のパラレルコーパスは、利用できるコーパスが少なく、獲得できる表現が限られてしまう。

Bannard & Callison-Burch (2005)は、英語とドイツ語の対訳コーパスを用いて、片方の言語（ドイツ語）をピヴォットの軸として、言い換え表現の獲得を行った。

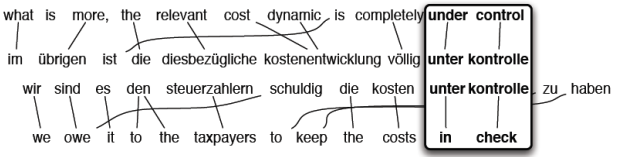


図1: Bannard & Callison-Burch (2005)の言い換え獲得

ドイツ語の表現(i.e., “unter kontrolle”)を軸に、その表現に訳されている英語の複数の表現(i.e., “under control” と”in check”)を言い換え候補として獲得した。さらに文脈内で最尤の言い換え候補を選出するため、その文内において最もNgram確率が高い候補を言い換え表現として獲得した。

## 3　提案手法

Bannard & Callison-Burch (2005)の方法を用いることで、「他の言語で同一の単語に訳される表現は同義と成りうる」という特徴を使って同義述語の獲得が可能である。しかし、本稿で対象とするのは”purchase”といった述語であるため、単純なNgram確率よりも、「項」の情報が重要な手がかりになると考えられる。そこで、本稿では、対訳コーパスを用いて獲得された同義述語候補に対し、述語の項を素性とした類似度を加えることで、同義述語獲得の精度向上を目指す。提案手法の処理は下記のとおりである。なお、本研究の対象である英語を E(English)、対訳コーパスの外国語をF(Foreign)とする。

i.　対訳コーパスから自動単語アライメントを用いて2種類の翻訳辞書を獲得する。1つ目はEを原言語(Source)、Fを翻訳先言語(Target)として作成した翻訳辞書である(E→F)。もう1つはその逆(F→E)である。
ii.　iで作成した2種類の翻訳辞書をもとに、Fで同一の表現に翻訳されているEの単語ペアの獲得を行い、これらを同義

述語候補とする。

iii. 英文コーパスより項を素性とした述語の類似度モデルを作成し、ii で獲得した同義述語候補ペアの類似度を計算する。

iv. 翻訳確率と類似度を組み合わせ、スコアの高いものを同義述語として獲得する。

### 3.1 翻訳辞書を用いた同義述語候補の獲得

対訳コーパスを用いて、同義述語候補の獲得を行う。本稿では、内山・井佐原(2002)の「日英新聞記事対応付け」データを用いた(日英対訳コーパス, 15 万文)。

対訳コーパスに対して、giza++(Och & Ney, 2003)を用いて翻訳辞書を作成した。なお、アライメントの精度をよくするため、英語の Head を後ろに移動させて日本語と同様の語順にし、英語の冠詞や日本語の格助詞を削除した(e.g., Isozaki et al., 2010)。Parser は The Stanford Parser(Klein and Manning, 2003)を用いた。下記が、獲得された翻訳辞書の例である。

| Source(E) | Target(F) | 翻訳確率 |
|---|---|---|
| buy | 購入 | 0.289 |
| purchase | 購入 | 0.512 |
| buy | 買う | 0.591 |
| purchase | 買う | 0.181 |

図 2: 翻訳辞書の例（Source→英語、Target→日本語）

| Source(F) | Target(E) | 翻訳確率 |
|---|---|---|
| 買う | purchase | 0.063 |
| 購入 | purchase | 0.516 |
| 買う | buy | 0.228 |
| 購入 | buy | 0.349 |

図 3: 翻訳辞書の例（Source→日本語、Target→英語）

これらの翻訳辞書と翻訳確率をもとに、F（i.e., 日本語)を軸に、同義述語候補の獲得を行った。なお、本稿では述語を対象としているので、WordNet に記載されている動詞と形容詞のみを対象にした。同義述語候補の獲得は、Bannard & Callison-Burch (2005)と同様に、下記の式をもとに、英語の述語である $e_1$ と $e_2$ に対し、同じ外国語の単語($f_1$..$f_n$)に訳された場合の翻訳確率を総和して求める。このスコアを、TP-Score(Scores based on Translation Probability)と本稿では呼ぶ。図 4 は、獲得された同義述語候補の例である。

(5) $$p(e_2|e_1) = \sum_f p(e_1|f)(f|e_2)$$

| $e_1$ | $e_2$ | TP-Score |
|---|---|---|
| buy | purchase | 0.224 |
| buy | spree | 0.199 |
| buy | acquire | 0.153 |

図 4: 獲得された同義述語候補の例

例が示すように、同義述語である”buy”と”purchase”が高い TP-Score を出しているが、同義ではない”buy”と”spree”というペアにも高い確率が付与されている。これは、”buying spree(買いまくる)”という表現が頻出するため、誤って”spree”と“買う”がアライメントされてしまったことによるものだと考えられる。[1]

[1] “spree”は非常に稀ではあるが、WordNet には動詞としての用法が記載されているため、同義候補の対象とした。

### 3.2 項を素性とした述語の類似度

3.1 で示したように、対訳コーパスのみから同義述語候補を獲得すると、アライメントエラーの影響で、まったく意味の異なる述語同士が高いスコアで獲得されてしまう。これらを排除するために、英語のコーパス単独から項を素性として計算した述語の類似度を加える。

類似度の計算には、2010 年 4 月～5 月にクロールした英文ブログデータ約 200 万文と、Wall Street Journal から約 5 万文、American National Corpus から約 20 万文を抽出して用いた。これらデータから、The Stanford Parser の Typed Dependency 情報をもとに、「項-項の種類-述語(e.g., government-nsubj-buy)」の 3 つ組を抽出した。類似度は、Lin(1998)に基づき下記の通りに計算した。以後、この述部類似度を similarity score という意味で SIM-Score と呼ぶ。f は素性（すなわち、「項-項の種類」）を表す。I(S)は、素性セット S に含まれる情報量を表す。

(6) $$\mathrm{sim}(e_1, e_2) = \frac{2 \times I(F(e_1) \cap F(e_2))}{I(F(e_1)) + I(F(e_2))}$$

(7) $$\mathrm{I(S)} = -\sum_{f \in S} \log P(f)$$

(6)の分子は、$e_1$ と $e_2$ が共通して持つ素性の情報量に 2 を積算したものである。例えば、図 5 のような素性が抽出された場合、”stock-dobj”, “government-nsubj”, “property-dobj”が $e_1$（すなわち"buy"）と $e_2$（すなわち”purchase”）が共通して持つ素性である。分子は、それら素性のそれぞれの情報量を総和した値になる。分母は、”buy”の素性すべての情報量と”purchase”の素性すべての情報量を総和した値である。

| $f$ | buy | purchase |
|---|---|---|
| stock-dobj | x | x |
| votes-dobj | x | |
| government-nsubj | x | x |
| property-dobj | x | x |

図 5: 素性の例(“x”は素性との共起を表す)

### 3.3 TP-Score と Sim-Score の統合

3.1 と 3.2 で計算したスコアを統合する。本稿では、下記の式でスコアを統合した。例を図 6 に示す。

(8) 同義述語スコア = TP-Score * SIM-Score

| $e_1$ | $e_2$ | TP-Score | SIM-Score | TP-Score* SIM-Score |
|---|---|---|---|---|
| buy | purchase | 0.224 | 0.215 | **0.04861** |
| buy | spree | 0.199 | 0.0003 | **0.00006** |
| buy | acquire | 0.008 | 0.131 | **0.00105** |

図 6: TP-Score と SIM-Score を統合した例

図 6 が示すように、アライメントエラーによって高い TP-Score をもった”buy”と”spree”のペアに対しては、項の類似度を取り入れることで、スコアを下げることが出来た。

上記を用いて獲得された同義述語候補から、同義述語スコア 0.001 以上でかつ上位 10%のものを「同義述語」として獲得する。

表 1: 実験 1 結果（日英コーパス）

| | BL1: TP-only | BL2: TP + Ngram | Proposed: TP + SIM |
|---|---|---|---|
| Prec | 70.2% (139/198) | 58.3% (7/12) | **87.3%** (89/102) |
| Rec | **49.3%** (139/282) | 2.5% (7/282) | 31.6% (89/282) |
| F | **0.58** | 0.05 | 0.46 |

表 2: 実験 2 結果（英仏コーパス）

| | BL1: TP-only | Proposed: TP + SIM |
|---|---|---|
| Prec | 60.4% (229/379) | 69.5% (189/272) |
| Rec | 81.2% (229/282) | 67.0% (189/282) |
| F | 0.69 | **0.68** |

表 3: 辞書をマージした結果

| | Merged |
|---|---|
| Prec | **89.2%** (83/93) |
| Rec | 29.4% (83/282) |
| F | 0.44 |

## 4 実験・評価

### 4.1 評価データ

人手で同義述語の評価データを作成した。データは、3.2 で使用したブログと同様のものを用いて、ランダムに述語を抽出し作成した。Synonym（同義）、Antonym（反義）、Others（その他）の 3 種類のデータを作成した。データは著者の一人が作成した。評価は、Synonym を正例、Antonym と Others を負例として行った。

| Relation | Total | Examples |
|---|---|---|
| Synonym | 282 | buy v.s. purchase |
| Antonym | 78 | buy v.s. sell |
| Others | 405 | buy v.s. read |

図 7: 評価データの内訳と例

### 4.2 実験 1 と結果（日英コーパス）

内山・井佐原(2002)の「日英新聞記事対応付け」データを用いて提案手法を評価した。比較手法として、TP-Score のみを用いたものと、Bannard & Callison-Burch (2005)の、TP-Score に Ngram 確率を加えたものを用いた。Bannard & Callison-Burch (2005)は特定の文脈において、$e_1$ の同義表現の候補集合($e_2..e_n$)を $e_1$ と置換してもっとも確率の高い表現を獲得し評価した。本稿では、同義述語辞書の獲得が目的であるため、各英語の述語 $e_1$ に対し、コーパス内で $e_1$ が出現した文を、同義述語候補である $e_2..e_n$ に置換し、Ngram スコアの上位 10 件の $e_n$ を同義述語として獲得した。[2]

- **Baseline(BL)1:TP-only**
  TP-Score のみで獲得した同義述語辞書(0.001 以上かつ上位 10%)
- **Baseline(BL)2:TP + Ngram**
  TP-Score に Ngram スコアを加えて獲得した同義述語辞書(上位 10 件)
- **Proposed:TP + SIM**
  TP-Score に SIM-Score を加えて獲得した同義述語辞書(0.001 以上かつ上位 10%)

評価は Precision(Prec)、Recall(Rec)、F 値(F)を用いて行う。実験の結果を表 1 に示す。

$$\text{Precision} = \frac{\text{正解の Synonym} \cap \text{システムが獲得した同義述語}}{\text{システムが獲得した同義述語}}$$

$$\text{Recall} = \frac{\text{正解の Synonym} \cap \text{システムが獲得した同義述語}}{\text{正解の Synonym}}$$

$$\text{F 値} = \frac{2*Precision*Recall}{Precition+Recall}$$

### 4.3 実験 2 と結果（英仏データ）

日英コーパスを用いた場合、どの手法も F 値が比較的低かったため、英語と語順が類似しているフランス語との対訳コーパスを用いて、再度、提案手法の評価を行った。コーパスは Europarl parallel corpus(Koehn, 2005)の英仏対訳コーパス(200 万文)を用いた。結果を表 2 に示す。表 2 が示すように、日英コーパスに比べ、英仏コーパスを用いた方が、F 値が向上した。さらに、4.2 の日英対訳データと 4.3 の英仏対訳データから獲得した同義述語辞書をマージし、双方の辞書に出現したもののみを獲得して評価した。[3]結果を表 3 に示す(Merged)。複数の対訳コーパスから作成した辞書のマージを取ることで、Precision が最も高くなった。

## 5 考察・結論

「他の言語で同一の単語に訳されている表現は同義と成りうる」という特徴を検証すべく、対訳コーパスを用いて英語の同義述語の獲得を行った。日本語とフランス語をピヴォットとして、同じ単語に訳されている異なる英語の述語を獲得したところ、同義の述語を獲得することが出来た。

| $e_1$ | $e_2$ |
|---|---|
| buy | purchase |
| emphasize | stress |
| sack | dismiss |
| tap | bug |

図 8：獲得された同義述語の例

“buy”と”purchase”のような典型的な同義述語だけではなく、”sack”と”dismiss” (He got sacked/dismissed from his job.)や、”tap”と”bug” (The phone line has been tapped/bugged.)といった、WordNet などのシソーラスにはエントリのない同義述語を獲得することができ、対訳コーパスを用いて同義述語を獲得する可能性を確認できた。

しかし、対訳コーパスのみから同義述語を獲得する場合、アライメントエラーの影響で、誤った単語対も同義述語として獲得してしまった。

これらのアライメントエラーによる影響を軽減するため、英文コーパスから学習した述語の類似度を加えた。実験の結果、F 値ではどの手法も同等の精度となったが、対訳コーパスからの TP-Score のみを用いる場合と、類似度を組み合わせた場合、また

[2] コーパスは 3.1 の対訳新聞コーパスの英文を利用した。

[3] 閾値 0.001 以上の候補をすべて用いた。

さらに複数の対訳コーパスからの結果を組み合わせた場合では、Precision と Recall に違いが出た。

**・異なる手法の組み合わせ**

対訳コーパスのみを用いて同義述語候補を獲得した場合(BL1)、アライメントエラーによって、”buy”と”spree”の様な誤った同義述語候補も獲得された。結果として、Precision が最も低い値となった。

| $e_1$ | $e_2$ |
|---|---|
| buy | spree |
| issue | discount |
| water | clean |

図 9: TP-Score によって誤って抽出された同義述語候補

一方、英語コーパスから計算した述語の類似度をさらに組み合わせることで、図 9 のような単語ペアを正し排除し、Precision をあげることが出来た。しかし、本稿の実験では類似度計算用のコーパスサイズが小さかったため、Recall を下げてしまった。今後は、類似度計算のコーパスを増やして再度考察したい。

Bannard & Callison-Burch (2005)の Ngram 確率を合わせた手法(BL2)は、本稿が目的とする同義述語の獲得では有益ではなかった。これは、Bannard & Callison-Burch (2005)はフレーズ(e.g., “in check”)を単位に言い換え表現の獲得を行っていたが、本稿では単語一語(e.g., “buy”)を対象としているからであると考えられる。単語のみを置換して Ngram スコアを計算する方法では、述語同士の同義性を判定することが難しい（たとえば、”I bought a smartphone.”も”I want a smartphone.”も、どちらも自然である）。そのため、Ngram 確率を用いると精度が下がってしまったと考えられる。

また、日英コーパスと英仏コーパスの 2 つのコーパスから獲得された同義述語候補のマージを取ると、Precision が向上した（表 3）。複数の対訳コーパスを組み合わせることは効果があると言える。

実験結果が表すように、対訳コーパスから単独に同義述語を獲得するよりも、類似度を合わせたり、さらに複数の言語での結果を組み合わせた方が高い Precision で同義述語を獲得することが出来る。また、実験では、単純に上位 10%の候補を獲得する方法をとったが、単語によってはそもそも同義の単語が存在しづらい(e.g. 「水をあげる」という意味の”water”)ものがあり、どの候補を獲得するかは今後の課題とする。

**・日英コーパスと英仏コーパスの違い**

日英コーパスと英仏コーパスの 2 つを用いて同義述語を獲得した。上位 10%を抽出して評価した結果では、日英コーパスの方が Precision が高く、英仏コーパスの方が Recall が高かった。より詳しく違いを考察するために、同義述語スコアの閾値を 0.05～0.0005 に変化させて考察した。

図 4 が示すように、英仏コーパスの方がどの閾値でも獲得された同義述語辞書の F 値が高く、Precision も閾値をあげることで向上した。これらの違いは、アライメントエラーの影響によるものであると考える。英仏コーパスの方が、コーパスサイズが大きく、また、フランス語と英語の方が語順などが比較的似ているため、アライメントの精度が高かった。一方、日英コーパスからは、”attempt”と”murder”、”start”と”talk”などが同義述語として誤って獲得されてしまった。これらは”XX attempted murder”や”start to talk”などの構造と日本語の述部構造のアライメントがうまく取れなかった結果であると推測できる。また、今回の実験からでは、獲得された同義述語そのものには、コーパス間で大きな差は見られなかった。今後は、コーパスを増やし、ピヴォットとなる言語の違いによって獲得される同義述語の傾向が異なるかどうかを詳しく考察したい。

表 4 日英・英仏コーパスの結果比較

| 閾値 | 日英コーパス | | | 英仏コーパス | | |
|---|---|---|---|---|---|---|
| | Prec | Rec | F | Prec | Rec | F |
| 0.05 | 100.0% | 0.7% | 0.01 | 91.7% | 7.8% | 0.14 |
| 0.01 | 92.9% | 9.2% | 0.17 | 89.0% | 25.9% | 0.40 |
| 0.005 | 93.2% | 14.5% | 0.25 | 82.8% | 39.4% | 0.53 |
| 0.001 | 86.9% | 30.5% | 0.45 | 70.4% | 61.7% | 0.66 |
| 0.0005 | 81.8% | 38.3% | 0.52 | 66.9% | 68.8% | **0.68** |
| 0.0001 | 67.9% | 52.5% | 0.59 | 57.2% | 78.7% | 0.66 |
| 0.00005 | 67.9% | 53.2% | **0.60** | 56.3% | 80.9% | 0.66 |

本稿では、対訳コーパスを用いて英語の同義述語の獲得を検討した。日英・英仏対訳コーパスを用いて、同一の単語に訳されている英語のペアを獲得したところ、それらは同義である可能性が高いことが分かった。しかし、アライメントエラーの影響で誤りを含む同義述語候補が獲得されたため、項を素性とした類似度を加えることで獲得した同義述語辞書の Precision を向上させた。今後は、コーパスの種類やサイズを増やして実験を行うとともに、単語一語だけではなく、”use up”などのフレーズを対象にした同義述語表現の獲得を検討したい。

## Acknowledgment

本研究を進めるにあたり、貴重なご意見をくださった京都大学大学院情報学研究科　中澤 敏明先生、NTT コミュニケーション科学基礎研究所　須藤 克仁研究員に深く御礼申し上げます。

## References

Bannard C., and Callison-Burch C. (2005). Paraphrasing with bilingual parallel corpora. *Proceedings of the 43rd Annual Meeting of the ACL,* 597-604.

Barzilay, R., and McKeown, K, R. (2001). Extracting paraphrases from a bilingual corpus. *Proceedings of the 39th Annual Meeting of the ACL,* 50-57.

Shinyama, S., Sekine, S., and Sudo, K. (2002). Automatic paraphrase acquisition from news articles. *Proceedings of the 2nd international conrerence on Human Language Technology Research,* 313-138.

Och, F. J., and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics, 29(1),* 19-51.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. *Proceedings of the Fifteenth International Conference on Machine Learning,* 296-394.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *Proceedings of the 10th Machine Translation Summit,* 79-86.

内山将夫・井佐原均 (2002). 日英新聞記事の対応付けと精度評価. *情報処理学会研究報告 2002-NL-151*, 15-22.

Klein, D., and Manning, C. D. (2003). Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting of the ACL,* 423-430.

Isozaki, H., Sudoh, K., Tsukada, H., and Duh, K. (2010). Head finalization; A simple reordering rule for SOV languages. *Proceedings of 5th Workshop on Statistical Machine Translation,* 250-257.