対訳コーパスからの訳語対抽出
– 多言語シソーラス自動構築に向けて–

辻　慶太 (東京大学大学院教育学研究科)
影浦　峡 (学術情報センター研究開発部)
芳鐘冬樹 (東京大学大学院教育学研究科)

Abstract

多言語シソーラスにおける同等関係自動構築という観点から，対訳コーパスからの訳語対自動抽出研究の概略をまとめた。次に緩やかな対訳コーパスから，出現・共起頻度に基づいて専門用語の対訳語構成要素対を抽出する場合に，コーパスからまず専門用語の訳語対候補を抽出し，これをアラインメントとして頻度を数える有効性を示した。訳語対候補の抽出条件としては，(1) 正解訳語対と品詞パターンが同じこと，(2) 専門用語辞書に含まれる対訳語構成要素対を 1 つ以上含むこと，の 2 つを用いた。さらに正解訳語対における語構成要素の対応パターンも援用した場合，抽出結果がさらに向上することを示した。

1　はじめに

シソーラスの利用は，情報検索に一定のパフォーマンス向上をもたらす。また多言語シソーラス中の同等関係 (2 言語間で等価なディスクリプタの関係) は，近年注目されている言語横断検索に有益な情報を提供する。だがシソーラスは，構築にかかるコストが常に問題となってきた。そこで現在筆者らは，既存の情報源を利用した，多言語シソーラス自動構築手法の開発を目指している。

多言語シソーラス中の同等関係構築には訳語対の把握が必須であり，何らかの情報源からの自動抽出が望まれる。ところで近年，計算機処理能力の向上や出版事情の変化に伴い，利用可能なコーパス，さらに対訳コーパスが増えてきている。コーパスとは大規模な電子テキストで，対訳コーパスとは同じ内容が 2 言語で書かれたコーパスである。対訳コーパスとしては英仏語で書かれたカナダ議会の議事録 : Hansard コーパスなどが有名である。コーパスには，辞書などに比べれば更新の早いものが多く，最新の用語が含まれている点で，多言語シソーラス自動構築に有用な情報源と考えられる。

自然言語処理分野では，対訳コーパスからの訳語対抽出研究が盛んに行われており，同等関係自動構築の参考になる。以下では，それら対訳コーパスからの訳語対抽出手法を概観・整理する。次に，統計的な訳語対抽出手法におけるアラインメントの問題を指摘し，改良に向けた簡単な実験結果を示したい。

2　訳語対抽出研究概観

対訳コーパスからの訳語対抽出手法は，語がコーパス中で出現・共起する頻度といった統計的情報に基づく手法と，音形の類似，辞書中の語あるいは意味カテゴリーの一致といった言語的情報に基づく手法の 2 つに大別できる。以下では，まず言語的情報に基づく抽出手法についてまとめ，次に統計的情報に基づく抽出手法についてまとめたい。

2.1　言語的情報に基づく抽出手法

まず音形の類似という観点から，訳語対抽出を図った研究としては，松尾 & 白井 (1996)，Collier ら (1997)，石本 & 長尾 (1994) がある。前者は対訳コーパス中の訳語候補の発音を推定し，それが一致する語同士を訳語対としている。中後者は対訳コーパスの日本語側に含まれるカタカナ語を英語に字訳し，字訳結果に類似した英語を，コーパスから抽出して訳語対としている。

山本 & 坂本 (1993)，石本 & 長尾 (1994) は，候補語の語構成要素に対して，辞書が挙げるすべての訳語の組み合わせを考え，コーパス中にその組み合わせと類似したものがあれば，それらを訳語対として抽出する手法を提案している。また Ker & Chang (1997) は，各語に意味カテゴリーを付与している辞書を 2 言語それぞれで入手し，両言語における意味カテゴリー同士の類似度を測った上で，類似した意味カテゴリーの語をコーパスから抽出することを提案している。

言語的情報に基づく手法は，対訳コーパス中の出現頻度が低い語でも抽出できる長所があるが，抽出できる語は，元々音形的に類似した語や，辞書に語構成要素・類似語が載っている語に限定される弱点がある。それに対して，次節で述べる統計的情報に基づく手法は，低頻度語の抽出に弱いものの，音形の似ていない語，辞書に類似語がない新語も抽出できる長所がある。

2.2　統計的情報に基づく抽出手法

ここでは統計的情報に基づく抽出手法についてまとめる。対訳コーパス中の対応する文や単語列で，語の出現・共起頻度を数える単位にされたものを，以下ではアライ

ンメントと呼ぶ。例えば，言語 J，E の対訳コーパスにおいて，語 $w_j(\in J)$ は 2 ，4 ，5 番目の訳文に出現し，語 $w_e(\in E)$ は 1 ，2 ，5 番目の訳文に出現したとする。このとき文をアラインメントに取るならば，語 w_j，w_e は 2 ，5 番目のアラインメントで共起し，共起頻度は 2 ，などと考える。

アラインメントに共起する度合いが高い語を，訳語対として抽出する統計的手法としては，Gale & Church (1991) ,van der Eijk (1993) ,Kupiec (1993) ,熊野 & 平川 (1994) , Fung (1995) , Melamed (1996) , Smadja ら (1996) , Vogel ら (1996) , 高尾 (1996) , 高尾ら (1996) , 高尾 (1997) , 北村 & 松本 (1997) , 佐藤 & 中西 (1997) , Ahrenberg ら (1998) , がある。以下ではこれらの研究で用いられている共起尺度についてまとめ，次にアラインメントの取り方についてまとめる。

2.2.1 共起尺度

ここで，出現・共起アラインメント数を表 1 のように定義する。w_j，w_e をそれぞれの言語の単位とすると，f_{11} は w_j と w_e が共起したアラインメント数，f_{00} は w_j も w_e も現れなかったアラインメント数，f_{10} は w_j が現れ，かつ w_e が現れなかったアラインメント数を表す。f_{01} はその逆である。また $f_{..}$ はアラインメントの総数を表す。

	w_j+	w_j-	Total
w_e+	f_{11}	f_{01}	$f_{.1}$
w_e-	f_{10}	f_{00}	$f_{.0}$
Total	$f_{1.}$	$f_{0.}$	$f_{..}$

表 1: w_j，w_e の出現・共起アラインメント数

まず Fung (1995) が抽出基準の中心に据え，Gale & Church (1991) , van der Eijk (1993) , 北村 & 松本 (1997) , Smadja ら (1996) が参考にした，特殊相互情報量 MI は次のように定義される[1] 。

$$MI = \log \frac{\frac{f_{11}}{f_{..}}}{\frac{f_{1.}}{f_{..}}\frac{f_{.1}}{f_{..}}}$$

Smadja ら (1996) , 高尾ら (1996) が抽出基準の中心に据え，北村 & 松本 (1997) が参考にした Dice 係数 DC は次のように定義される。

$$DC = \frac{f_{11} \times 2}{f_{1.} + f_{.1}}$$

北村 & 松本 (1997) は上記の Dice 係数が絶対的な共起アラインメント数を反映しない形になっていることを指摘し，共起アラインメント数の対数で重み付けした以下のような係数 KD を提案している。

$$KD = \log f_{11} \frac{f_{11} \times 2}{f_{1.} + f_{.1}}$$

この他には，Gale & Church (1991) の ϕ^2 統計量，Smadja ら (1996) が参考にした相互情報量，Fung (1995) が特殊相互情報量と共に抽出基準とし，また Ahrenberg ら (1998) が抽出基準とした t スコアなどがある。

さて，上記の統計的尺度はいずれも再帰計算を必要としない比較的単純なタイプである。それに対して，Kupiec (1993) , Melamed (1996) , 佐藤 & 中西 (1997) , Vogel ら (1996) の尺度のように，一定の値に収束するまで繰り返し計算を行う，より複雑なタイプもある。

2.2.2 アラインメントの単位

アラインメントの取り方は，用いる対訳コーパスの対訳性と密接に関わっている。文のほとんどが 1 対 1 に対応した対訳コーパスを用いた研究では，文がそのままアラインメントにされる場合が多い(Gale & Church (1991) , van der Eijk (1993) , Melamed (1996) , Smadja ら (1996) , Vogel ら (1996) , 高尾 (1996) , 高尾ら (1996) , 高尾 (1997) , 北村 & 松本 (1997) , 佐藤 & 中西 (1997) , Ahrenberg ら (1998))。

一方，文単位の対応がない緩やかな対訳コーパスから，出現・共起頻度によって訳語対を抽出する研究では，アラインメントの決定方法が 1 つのテーマとなる。熊野 & 平川 (1994) は，文単位の対応がない日英特許文書から，辞書が挙げる訳語対を高い割合で含む文同士をアラインメントとして抽出している。また Fung (1995) は，英中語の緩やかな対訳コーパスから，出現位置の近い語を DTW(Dynamic Time Warping) によって決定し，最初の訳語対とする。次にそれらの共起位置でコーパスを切ってアラインメントとし，残りの訳語対については，このアラインメントでの出現・共起頻度に基づいて抽出している[2] 。

文単位で対応した対訳コーパスを用い，文をそのままアラインメントとした研究は非常に多いが，文をアラインメントとすることがどの程度有効なのか，という考察はほとんど行われていない。また文をアラインメントにするならば，援用を検討すべき文固有の情報，係り受けのような内部構造を援用した研究もない。Fung (1995) においても同様で，この場合はアラインメントの言語学的意味も分かりにくい。

このような見地から現在，対訳コーパス中でのアラインメントの取り方，及びその内部構造の利用方法について検討している。次章の実験はその中間報告である。

3 実験

先述のように，言語的情報と統計的情報はそれぞれに長短所がある。言語の複雑な性質を考えれば，1 つの尺度で対訳コーパス中の訳語対がすべて抽出できるとは考えにくく，両方の情報を用いることが望ましい。

さて，文が 1 対 1 に対応した対訳コーパスはあまり一般的でないが，抄録や記事といったレベルで対応した対訳コーパスは，比較的入手しやすい。情報源が現実に確保できる多言語シソーラス自動構築手法を目指すならば，文レベルに対応したコーパスではなく，緩やかな対訳コーパスを前提とした方が良い。

以上の 2 点から，抄録単位で一定の対応がある学会発表の日英抄録をコーパスとし，統計的情報も用いた訳語対抽出を今後図る予定である。このとき，抄録をそのま

[1] Fung (1995) , Gale & Church (1991) , van der Eijk (1993) , 北村 & 松本 (1997) は，これを相互情報量 (mutual information) と呼んでいるが，本研究では Smadja ら (1996) の区別に従って特殊相互情報量 (specific mutual information) と呼ぶことにする。

[2] DTW ですべての訳語対を抽出しないのは，DTW の計算コストが大きいからである

まアラインメントにすると，計算量及びノイズの多さが問題になるため，より小さく適切なアラインメントを設定したい。こうした観点，及び先ほどの先行研究の不足を補う観点から，アラインメントの設定方法，及びその内部構造の利用法について考えてみた。

3.1 手法

以下では，専門用語の訳語対を対訳用語対と呼ぶ。同等関係の構築にはこの対訳用語対が必要である。また対訳用語対中で対応する語構成要素の最小の組み合わせを対訳単位対と呼ぶ。例えば，< 射影特異点, singular point of projection> は対訳用語対，そこでの < 射影,projection> ，< 特異,singular> ，< 点,point> は対訳単位対である。さて対訳単位対は，一般に対訳用語対より出現頻度が高く，統計的尺度が適切に機能しやすい。そこで，訳語対抽出の基本的流れとしては，統計的情報に基づいてまず対訳単位対を抽出し，次にそれらの組み合わせという言語的情報を併用して，対訳用語対を抽出することを考えている。今回は，その対訳単位対の抽出に適切なアラインメントの取り方を考えてみた。

まずアラインメントの取り方として次の 2 つを考えた。1 つは抄録をアラインメントとする取り方である。もう 1 つは「事前に調べたコーパス中の正解対訳用語対と品詞パターンが同じで，かつ専門用語辞書に含まれる対訳単位対 (以下，辞書対と呼ぶ) を 1 つ以上含む単語列」を各抄録から対訳用語対候補として抽出し (部分的な重なりも認めた)，これをアラインメントとする取り方である。後者の場合はさらに，そこに出現・共起した頻度をそのまま用いる場合と，語構成要素の対応パターンを考え，対応パターンにない組み合わせについては，そのアラインメント内で共起しなかった，と数える場合の 2 通りを実験した。

ここで品詞パターンとは，対訳用語対の日本語・英語の品詞列であり，対応パターンとは，対訳用語対の内部で対応している語構成要素の位置に関するパターンである。例えば，対訳用語対 < 射影 / 特異 / 点, singular point of projection> の‘射影’，‘特異’，‘点’の品詞がそれぞれ J_1 ，J_2 ，J_3 で，‘singular’，‘point’，‘of’，‘projection’の品詞がそれぞれ E_1 ，E_2 ，E_3 ，E_4 であったならば，この対訳用語対の品詞パターンは $< J_1, J_2, J_3 | E_1, E_2, E_3, E_4 >$ である。また語構成要素として，日本語の 1 番目の単位‘射影’と英語の 4 番目の単語‘projection’が対応していることなどから，$< 1,4 >< 2,1 >< 3,2 >$ が，品詞パターン $< J_1, J_2, J_3 | E_1, E_2, E_3, E_4 >$ 中の対応パターンとなる。当然 1 つの品詞パターンに複数の対応パターンが存在する場合もある。対応パターンを考えて共起頻度を数える場合，仮に‘図書館 / 設置 / 率’と‘establishment rate of library’が，上と同じ品詞パターンでコーパス中に現れ，かつ対訳用語対候補，即ちアラインメントとされた場合，‘図書館’と‘library’はこのアラインメントにおいて共起したと考えるが，‘図書館’と‘establishment’などは共起しなかったと考える。

対訳単位対は対訳用語対に含まれるのだから，対訳用語対の可能性がある単語列をアラインメントにしてはどうか，というのが基本的な発想である。またアラインメント内の語をすべて均一に扱うのではなく，対訳用語対における内部対応パターンまで利用して，共起を考える点が本手法の特徴である。これには共起ノイズを大幅に減少できる利点がある。また多言語シソーラスの構築では，専門用語以外の語を積極的に除外する必要があるかもしれないが，このように専門用語の可能性がある単語列だけをアラインメントとすることで，一般語の多くが抽出対象の外に出る利点がある。

本手法最大の問題点は，対訳用語対の品詞・対応パターンには，分野・データに依らない一定の普遍性があり，これを用いても抽出手法としての一般性は失われないとする立場を取っていることであろう。この点の検証は今後の課題としたい。

さて，アラインメントの取り方は以上とし，共起尺度としては，先述の訳語対抽出研究で一般的な，特殊相互情報量，Dice 係数，北村らの修正 Dice 係数の 3 つを用いた。

3.2 データ

実験ではまず，辞書対として『人工知能大辞典』の索引部分に挙げられている対訳用語対 3742 個に含まれる対訳単位対 3436 個のうち，前置詞を除いたもの 3426 個を用いた。対訳コーパスとしては，学術情報センターの学会発表データベースに含まれる人工知能分野の日英抄録 1767 個を用いた。この抄録は文レベルの対応はあまりないものの，各抄録全体は一応等価な内容が仮定できる。

抄録の日本語・英語部分はそれぞれ茶筌，Brill tagger で形態素解析を行った。茶筌が切り出した単位と英単語の観点から数えた場合，抄録における専門用語の対訳単位対は，異なりで 5128 個であった。そのうち茶筌のコード ‘6 1 0 0’ と ‘6 2 0 0’，及び Brill tagger のコード NN*，JJ*，VB*それぞれ 1 つずつから成る対訳単位対は 3438 対で，そこから辞書対を除くと 2431 対が残った。今回は予備実験としてこれらを抽出タスクとし，抽出精度・再現率を算出した。正解対訳用語対の品詞・対応パターンは 1207 種類であった。

3.3 結果

特殊相互情報量，Dice 係数，北村らの修正 Dice 係数による抽出結果は，それぞれ図 1，2，3 のようになった。いずれの図でも曲線は，下から順に抄録をアラインメントとした場合，対訳用語対候補をアラインメントとした場合，さらに対応パターンまで考慮した場合，の精度・再現率を表している。辞書対を 1 つ以上含むという条件を加えたため，すべての正解対訳用語対がアラインメントになった訳ではなく，従ってすべての対訳単位対がアラインメントに含まれた訳ではない。そのため再現率の上限は，抄録をアラインメントとした場合に譲るものの，抽出精度の向上は，それを補って余りあるように思われる。

4 おわりに

本研究では，対訳コーパスからの訳語対抽出研究に関する概略・問題点をまとめ，その上で，出現・共起頻度に基づいて対訳単位対を抽出する場合に，対訳用語対候補をアラインメントとする有効性，及び内部の対応パターンを援用することの有効性を示した。

本研究は，多言語シソーラス自動構築に向けた，広範な実験のごく一部であり，今後，本研究結果の検証と，それを踏まえた自動構築手法を提案していく予定である。今後の課題としては，まずアラインメント抽出に有効か

つ一般性のある品詞・対応パターンの特定，同様に，有効な辞書対の質・量の検証，さらに抽出した対訳単位対を，辞書対に加えて再帰的に対訳単位対抽出を行う有効性の検証，などが考えられる。

参考文献

[1] Ahrenberg, L., Andersson, M. and Merkel, M. (1998) "A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts," *COLING'98*, p.29-35.

[2] Collier, N., Kumano, A. and Hirakawa, H. (1997) "Acquisition of English-Japanese Proper Nouns from Noisy-Parallel Newswire Articles Using KATAKANA Matching," *Proceedings of the Natural Language Processing Pacific Rim Symposium 1997*, p.309-314.

[3] Fung, P. (1995) "A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora," *The 33rd Annual Conference of the ACL*, p.236-233.

[4] Gale, W. A. and Church, K. W. (1991) "Identifying Word Correspondences in Parallel Texts," *Proceedings, DARPA Speech and Natural Language Workshop*, p.152-157.

[5] Ker, S. J. and Chang, J. S. (1997) "A Class-Based Approach to Word Alignment," *Computational Linguistics*, vol.23, no.2, p.313-343.

[6] Kupiec, J. (1993) "An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora," *The 31st Annual Meeting of the ACL*, p.17-22.

[7] Melamed, I. D. (1996) "Automatic Construction of Clean Broad-Coverage Translation Lexicons," *2nd Conference of the Association for Machine Translation in the Americas*, p.125-134.

[8] Melamed, I. D. (1997) "Automatic Discovery of Non-Compositional Compounds in Parallel Data," *2nd Conference on Empirical Methods in Natural Language Processing*.

[9] Shapiro, S. C. and Eckroth, D. [編]; 大須賀節雄 [監訳] (1991)『人工知能大辞典』丸善. [Shapiro, S. C. and Eckroth, D. (eds.) (1987) *Encyclopedia of Artificial Intelligence*, New York: J.Wiley & Sons.]

[10] Smadja, F., McKeown, K. R. and Hatzivassiloglou, V. (1996) "Translating Collocations for Bilingual Lexicons: A Statistical Approach," *Computational Linguistics*, vol.22, no.1, p.1-38.

[11] van der Eijk, P. (1993) "Automating the Acquisition of Bilingual Terminology," *Proceedings, Sixth Conference of the European Chapter of the Association for Computational Linguistics*, p.113-119.

[12] Vogel, S., Ney, H. and Tillmann, C. (1996) "HMM-Based Word Alignment in Statistical Translation," *COLING'96*, p.836-841.

[13] 石本浩之・長尾真 (1994) "対訳文書を利用した専門用語対訳辞書の自動作成：訳語対応における両立不可能性を考慮した手法について," 自然言語処理, 102-11, p.81-88.

[14] 北村美穂子・松本裕治 (1997) "対訳コーパスを利用した対訳表現の自動抽出," 情報処理学会論文誌, vol.38, no.4, p.727-736.

[15] 熊野明・平川秀樹 (1994) "言語情報と統計情報を用いた対訳文書からの機械翻訳辞書作成," 自然言語処理, 100-12, p.89-96.

[16] 佐藤健吾・中西正和 (1997) "最大エントロピー法による対訳単語対の抽出," 自然言語処理, 122-4, p.21-27.

[17] 高尾哲康 (1996) "対訳テキストコーパスからの対訳複合語の自動抽出," 山形県立米沢女子短期大学紀要, no.31, p.31-42.

[18] 高尾哲康・富士秀・松井くにお (1996) "対訳テキストコーパスからの対訳語情報の自動抽出," 自然言語処理, 115-8, p.51-58.

[19] 高尾哲康 (1997) "対訳対応類似度に注目した対訳複合語の自動抽出," 山形県立米沢女子短期大学紀要, no.32, p.85-92.

[20] 松尾義博・白井諭 (1996) "発音情報を用いた訳語対の自動抽出," 自然言語処理, 116-15, p.101-106.

[21] 山本由紀雄・坂本仁 (1993) "対訳コーパスを用いた専門用語対訳辞書の作成," 自然言語処理, 94-12, p.85-92.

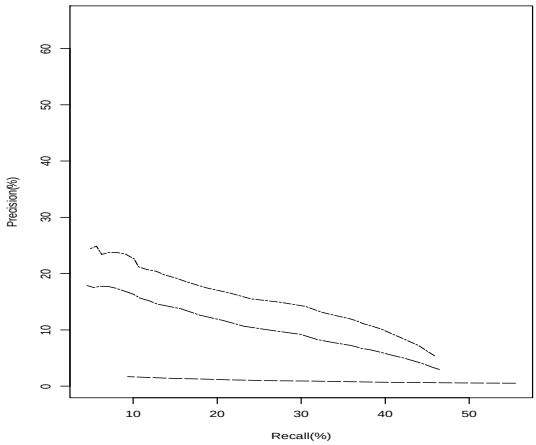

図 1: 特殊相互情報量

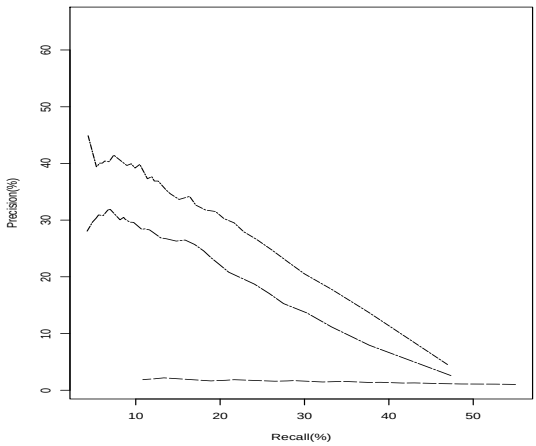

図 2: Dice 係数

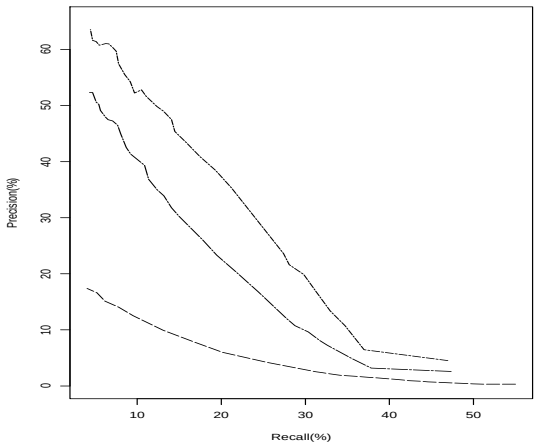

図 3: 北村らの修正 Dice 係数