対訳コーパスからの訳語対抽出における辞書情報の利用について

辻　慶太（東京大学大学院教育学研究科）
芳鐘冬樹（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）

Abstract

専門用語の語構成要素の訳語対（以下「対訳要素対」）は，コーパス中においてその周辺に，専門用語辞書が挙げる対訳要素対を持つ場合が多いことを，人工知能分野の日英抄録を用いてまず確認した。次に同抄録から，統計的尺度だけで対訳要素対を抽出する場合と，専門用語辞書が挙げる対訳要素対を近くに持つ対だけに絞り込む場合とで抽出結果を比較し，絞り込みが再現率をあまり落とさずに，精度を高めることを確認した。

1 はじめに

専門用語の語構成要素の訳語対（以後「対訳要素対」と呼ぶ）を，コンパラブルなコーパスから，辞書情報を援用して抽出する手法を検討する。

言語横断検索には専門用語の機械翻訳が1つの有効なアプローチと考えられる。その為の翻訳用辞書，あるいは多言語シソーラスをカレントな状態に維持するには，何らかのテキストコーパスから専門用語の訳語対（以下「対訳用語対」）を自動抽出するのが望ましい。対訳用語対抽出では，単位語に加え，複合語の対訳用語対を如何に抽出するかが問題となる。その場合，van der Eijk (1993)，大森ら (1997)，北村 &松本 (1997)のように，複合語を1言語内であらかじめ決定してから，それらを1つの単語のように扱って，訳語対抽出を図る方向もあるが，対訳用語対には，語構成要素同士が訳語対として対応するものも多い為，高尾ら (1996)のように，対訳要素対の情報も参考に抽出を図る方が効果的と思われる。そこで，対訳用語対の自動抽出に向けた足がかりとして，本研究ではコーパスからの対訳要素対抽出を試みたい。

さて，従来の訳語対抽出研究では，文レベルで対応した対訳コーパスをデータに用いることが多いが（Gale & Church (1991), van der Eijk (1993), Haruno et al.(1996), Melamed (1996), Smadja et al.(1996)），そのようなコーパスは，一般に入手が困難である。対訳性のない二言語のコーパスから訳語対を自動抽出することも可能だが（Rapp (1995)，田中 &岩崎 (1995)，大森ら (1997)，Fung & McKeown (1997)），計算量やパフォーマンスの点で問題がある。それに対して，文レベルでの対訳関係はないものの，文以上の単位間では一定の対応が見られる，同一事件に関する新聞記事に代表されるコンパラブルなテキストは，情報源として比較的入手しやすく，計算コストも低く抑えられ，応用可能性が高い。そこで本研究ではそのようなコーパスを抽出源とした。

辞書に含まれる情報を，訳語対抽出に役立てる方法には様々なものが考えられ，今後の検証が待たれる。今回はそれに向けた一研究として，専門用語辞書が挙げる対訳要素対（以後「辞書対」と呼ぶ）をアンカーポイントとし，コーパス中におけるそれら対の周辺で共起する語同士を，優先的に訳語対とすることの有効性を検証する。具体的には，統計的な共起尺度だけで抽出した場合と，それらが辞書に含まれる対訳要素対の周辺に現れているかで絞り込んだ場合とで結果の比較を行いたい。辞書対を用いて，このような絞り込みを行うのは，複合語の専門用語は，辞書に載っていない語構成要素だけで構成されることは少なく，辞書に載っている対訳要素対の周辺を探すことで，それが一部を構成する専門用語の他の対訳要素対を抽出できる可能性が高いと考えるからである。

以上，本研究の特徴としては，専門用語の対訳要素対を抽出対象としていること，1文同士が対訳関係にないコンパラブルなコーパスを用いていること，辞書中の対訳要素対のコーパス中での位置を用いて抽出精度の向上を図っていること，の3点が挙げられる。

2 データ

実験には，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。1抄録は日本語は平均4.89文，英語は平均5.71文から成る。これら抄録は，文同士で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。

人工知能抄録の日本語部分は茶筌1.51で単位分割を行い，英語部分はPorterアルゴリズムで語形を統一した。茶筌が切り出した単位（以後「単語」と呼ぶ）と

英単語の観点から数えた場合，人工知能抄録における専門用語の対訳要素対は，異なりで 4867個であった。そのうち日英 1単語ずつから成る対訳要素対は 2891対(59.40%)であった。今回は，予備的調査としてこれら 1単語同士の対訳要素対抽出を試み，残りの複数単語から成る対訳要素対に関する検証は今後の課題としたい。

対訳要素対抽出に援用する辞書データには『人工知能大辞典』(丸善, 1991)の日英対訳用語対 3742個に含まれる対訳要素対 3436個を用いた。対訳用語対から対訳要素対を自動抽出する手法には辻ら (1998)があり，比較的高い精度で抽出できるが，今回は自動抽出による誤りの影響を除く為，対訳用語対からの対訳要素対抽出はすべて手作業で行った。

3 実験

以下ではまず，抄録中の日英単語の対の周辺位置について定義する。次に，日英単語対が訳語対の場合とそうでない場合とで，周辺位置に辞書対が存在する割合がどれほど異なるかを検証する。その上で，抄録から統計的に訳語対抽出を行った場合と，周辺位置における辞書対存在割合による絞り込みを加えた場合とを比較し，後者の方が結果が良いことを示す。

まず抄録中の日英単語対 u_j, u_eの周辺位置 (l_j, l_e)を，u_jとの，あるいは u_eとの間の単語数に，後の場合は'+'，前の場合は'−'を加えて(日本語，英語)の順に表す。例えば，日英抄録にそれぞれ，

「 … /リダクション/の/完備/集合/は … 」
"... complete set of reduction..."

とあった場合(ただし'/'は茶筌の分割位置を表す)，「リダクション」と'reduction'という日英単語対の $(+2,-1)$の位置には，「集合」と'set'，あるいは「集合は」と'set'...等が存在することになる。この時，辞書対に「集合, =set」があった場合，「リダクション」と'reduction'は，$(+2,-1)$の位置に，辞書対を持つと考える。

このようにして，日英単語対の周辺位置ごとに，辞書対がどのような割合で存在するかが分かる。ここで (l_j, l_e)における辞書対存在割合とは，「辞書対が存在する (l_j, l_e)の数/ (l_j, l_e)の数」と定義する。例えば，1つの日本語抄録において「リダクション」が 3回，対応する英語抄録に'reduction'が 2回出現し，周辺位置 $(+2,-1)$が $3\times2=6$個すべて存在し，そのうちの 4個に辞書対が存在した場合，「リダクション」と'reduction'に関する $(+2,-1)$における辞書対存在割合は，$4\div6\times100(\%)=67(\%)$となる。

3.1 周辺位置における辞書対存在割合

以下では「訳語対は，周辺位置における辞書対の割合が非訳語対よりも高い」という仮定を検証する。調査対象の訳語対は，先述の対訳要素対 2891対である。非訳語対には，無作為抽出で得た，対訳関係にない日英単語対 2891対を用いた。これら 2つの単語対集合に関して，$(\pm0,\pm0)$ ~ $(\pm3,\pm3)$における辞書対存在割合の平均を調べたところ，表 1のようになった。縦軸・横軸はそれぞれ日本語・英語を表し，各欄における上段が訳語対，下段が非訳語対に関する辞書対存在割合である。

表 1から，$(+0,+0)$即ち直後の語同士が，辞書対である割合は，非訳語対では 0.79%であるのに対し，訳語対では 15.43%に上り，その差が際立って大きいこと，次いで $(-0,-0)$即ち直前の位置における差が大きいこと，また一般に辞書対存在割合は，訳語対の方が非訳語対よりも高く，その差は日英単語対から遠い位置ほど小さいこと，が分かる。

3.2 訳語対の抽出

前節で，周辺位置における辞書対存在割合は，一般に訳語対の方が非訳語対よりも高いことを見た。これを利用することで訳語対抽出の効率が上げられるかもしれない。以下では，訳語対抽出において一般的な，統計的共起尺度による訳語対抽出を行った場合と，それに辞書対存在割合を用いた絞り込みを加えた場合とで結果を比較し，精度がどれだけ向上するかを調べる。今回は統計的共起尺度として比較的一般的な，Dice係数(Smadjaら (1996), 高尾ら (1996), 北村 & 松本 (1997))を用いる。日英単語対 u_j, u_eに関する Dice係数 DCは次のように定義される：

$$DC=\frac{f_{11}\times2}{f_{1\cdot}+f_{\cdot1}}$$

ただし，f_{11}は u_jと u_eが共起した抄録数，$f_{1\cdot}$, $f_{\cdot1}$はそれぞれ u_j, u_eが出現した抄録数，である。

各抄録内の日本語単語に対して，その対応英語抄録中で Dice係数が最大になる英単語を決定し，対訳要素対として抽出した場合と，それらの中から $(\pm0,\pm0)$ ~ $(\pm n,\pm n)$の範囲内に辞書対を 1つでも持つ対だけを抽出した場合，とで結果を比較する(ただし $0<n<3$)。

Dice係数の閾値を 0.1毎に 0から 0.9まで変化させた場合の精度・再現率を図 1に示す。横軸が再現率，縦軸が精度である。一番下の曲線が，Dice係数で抽出しただけの場合で，残りの曲線は下から順に，n=3, n=2, n=1, n=0の結果である。対を絞り込めば再現率は必ず減少するのだが，nが大きい場合，それほど再現率を損なわずに精度を高くできること，nが小さいほど精度が高くなることが分かる。

さて先ほど表 1で，訳語対・非訳語対の周辺位置における辞書対存在割合は，$(+0,+0)$で，次に $(-0,-0)$で，差が最も大きくなることを見た。そこで，$(-0,-0)$, $(-0,+0)$, $(+0,-0)$, $(+0,+0)$の 4つの位置における辞書対存在割合が 0でないという条件で，訳語対を絞り込んでみた。結果は図 2の通りである。一番下の曲線が Dice係数のみの結果で，下から順に $(-0,+0)$, $(+0,-0)$, $(-0,-0)$, $(+0,+0)$の結果である。直後の位置に辞書対が存在する対を抽出することで，大きく精度を向上できること，ただし再現率は Dice係数のみに比べて落ちること，が分かる。

以下の 2節では，補足的な実験・集計を行う。

3.2.1 助詞・前置詞を数えない場合

日本語における複合語，例えば「情報検索」はしばしば「情報を検索する」のように，助詞などを伴って，いわゆる開いた形で文中に現れることがある。そこで文中の機能語を除いて元の複合語に近い形に機械的に還元すれば，複合語における語構成要素間の距離のばらつきが減少し，訳語対抽出の効率が良くなるのではないか，と考え，予備調査として今回，日本語では助詞，英語では前置詞をカウントせずに距離の測定を行ってみた。先ほど同様，ある範囲内の結果及び直前・直後に関する結果は，それぞれ図 3・4のようになった。助詞・前置詞も 1単語にカウントする，先ほどの方法に比べ，若干の精度向上が見られる。

3.2.2 出現頻度 1の訳語対の結果

さて，日英共に出現頻度が 1で共起頻度も 1という日英単語対は，Dice係数が 1となる。このような対には，その Dice係数の高さに拘わらず訳語対でないものが多く，そこから真の訳語対を抽出するのは，1つの重要な課題と考えられる。本研究手法が，このような低頻度の訳語対抽出にどれだけ有効かを検証する為，先ほどの抽出実験の結果に対して，出現・共起頻度 1の日英単語対に限った集計も行ってみた。表 2に結果を示す。各行で，上段は助詞・前置詞を 1単語にカウントする場合，下段はしない場合の結果である。

全般に精度は上がるものの再現率の落ち込みが顕著に見られた。低頻度の対訳要素対は，ほとんどの場合，辞書対のような一般的な対訳要素対と組み合わさって専門用語を構成し，コーパス中に存在していると予想していた。だが n=3までを対象範囲にしても再現率が 70%代にとどまるのは，低頻度の対訳要素対が，辞書対から離れて存在する場合が多いことを示している。

4 おわりに

本研究のようなコンパラブルなコーパスでは，訳語候補が多く存在するが，コーパス内の出現位置の周辺に辞書対が存在するかで絞り込みを行うことで，再現率をそれほど犠牲にせず，抽出精度を向上できること，日英共に直後に辞書対がある対を訳語対として優先的に抽出することで，特に精度が高まることが，本研究で示された。

本研究手法は，本質的には辞書対を用いたテキストアラインメントと考えられる。その意味で，先行研究におけるテキストアラインメント手法を適用して，抄録内をアラインメントに分け，その後に Dice係数に基づいて訳語対抽出を図るという方向も比較対象として考えられる。だが本手法は，候補対のごく周辺に辞書対があるかで絞り込みを行う為，あらかじめアラインメントを切り出した場合にも，そのアラインメントが文のようにある程度大きいものであれば，併せて適用できる手法である。従って，比較というよりも，併せて適用した場合にどの程度効果が上がるかを今後検証したい。また，本研究が取り上げなかった複数単語から成る対訳要素対に関する検証，辞書対までの距離で重み付けする訳語対抽出手法の検討も，今後行う予定である。

参考文献

[1] 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.]

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

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

[4] 田中久美子・岩崎英哉 (1995) "非対訳コーパスを用いた訳語関係の抽出." 自然言語処理, 110–13, p. 87–95.

[5] Rapp, Reinhard (1995) "Identifying Word Translations in Non-parallel Texts." *Proceedings of the 33rd Meeting of the Association for Computational Linguistics*, p. 320–322.

[6] 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.

[7] Haruno, M., Ikehara, S. and Yamazaki, T. (1996) "Learning Bilingual Collocations by Word Level Sorting." *COLING'96: The 16th International Conference on Computational Linguistics*, p. 525–530.

[8] 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.

[9] Smadja, F., McKeown, K. R. and Hatzivassiloglou, V. (1996) "Translating Collocations for Bilingual Lexicons: A Statistical Approach." *Computational Linguistics*, 22(1), p. 1–38.

[10] 大森久美子・佐藤健吾・中西正和 (1997) "共起関係を利用した対訳コーパスからの連語の対訳表現抽出." 自然言語処理, 122–3, p. 13–20.

[11] 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, 38(4), p. 727–736.

[12] Fung, P. and McKeown, K. (1997) "Finding Terminology Translations from Non-Parallel Corpora." *The 5th Annual Workshop on Very Large Corpora*, p. 192–202.

[13] 辻慶太・影浦峡・芳鐘冬樹 (1998) "隠れマルコフモデルに基づく日英専門用語対からの語構成要素対抽出：言語横断検索に向けて." 第 46回日本図書館情報学会研究大会発表要綱, p. 91–94.

$l_j \backslash l_e$	−3	−2	−1	−0	+0	+1	+2	+3
+3	1.00	1.31	1.38	1.14	1.09	0.97	0.94	0.84
	0.83	0.87	0.81	0.77	0.88	0.77	0.70	0.81
+2	1.17	1.28	1.35	1.31	1.04	1.24	1.20	0.85
	0.84	0.83	0.80	0.88	0.80	0.84	0.82	0.71
+1	1.38	1.63	2.32	2.01	2.77	3.42	0.99	0.86
	0.80	0.72	0.85	0.87	0.85	0.91	0.89	0.82
+0	1.40	1.88	2.48	1.80	15.43	1.08	1.00	1.22
	0.88	1.05	0.80	0.92	0.79	0.86	0.96	0.84
−0	1.10	1.12	1.85	8.98	2.46	1.90	1.35	1.25
	0.92	0.97	0.80	0.66	0.81	0.89	0.81	0.90
−1	0.79	1.12	2.44	1.62	2.30	1.49	1.15	1.12
	0.80	0.77	0.86	0.89	0.74	0.80	0.74	0.79
−2	0.78	1.25	1.29	0.92	1.64	1.07	1.14	1.19
	0.88	0.74	0.76	0.88	0.81	0.77	0.86	0.75
−3	1.07	0.83	0.92	0.99	1.75	0.83	1.16	1.15
	0.82	0.82	0.82	0.80	0.85	0.83	0.87	0.76

表 1: 位置 (l_j, l_e) に辞書対が存在する割合

	抽出対	正解対	再現率 (%)	精度 (%)
Baseline	11053	135	100.00	1.22
$n=0$	893	49	36.30	5.49
	420	38	28.15	9.05
$n=1$	1997	69	51.11	3.46
	956	60	44.44	6.28
$n=2$	3447	91	67.41	2.64
	1817	81	60.00	4.46
$n=3$	4718	101	74.81	2.14
	2639	96	71.11	3.64
$(+0,+0)$	304	30	22.22	9.87
	182	24	17.78	13.19
$(+0,-0)$	259	13	9.63	5.02
	103	7	5.19	6.80
$(-0,+0)$	198	2	1.48	1.01
	120	2	1.48	1.67
$(-0,-0)$	272	18	13.33	6.62
	97	10	7.41	10.31

表 2: 出現・共起頻度 1の対訳要素対の抽出結果

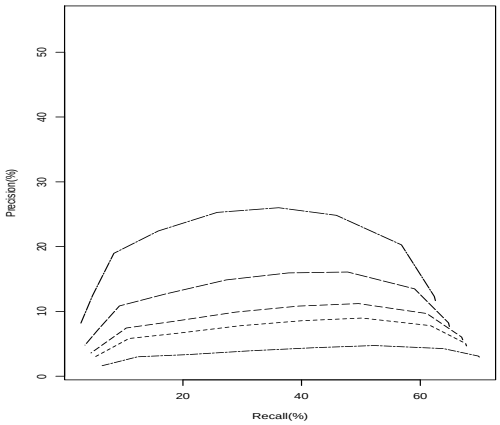

図 1: $(\pm n, \pm n)$ 以内に辞書対

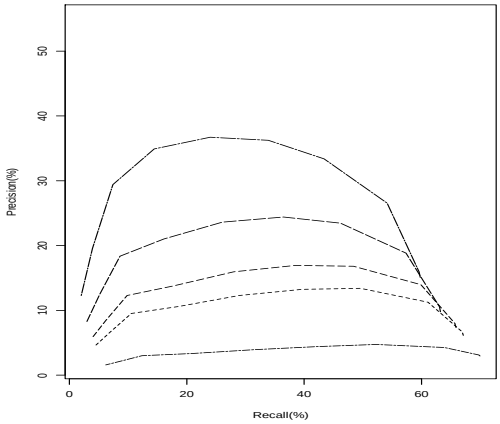

図 3: 助詞・前置詞を数えず $(\pm n, \pm n)$ 以内に辞書対

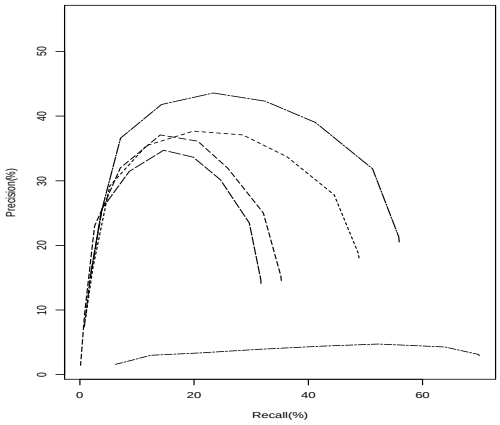

図 2: 直前・直後に辞書対

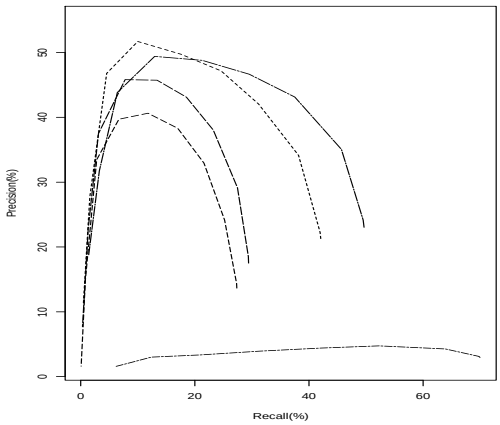

図 4: 助詞・前置詞を数えず直前・直後に辞書対