Vol.2015-MPS-105 No.2
2015/9/29

文構造情報を用いた対訳コーパスからの対訳辞書作成

新海 正吾[1,a)]　佐藤 大輔[1,b)]　松永 務[1,c)]

概要：情報資産の有効活用の動きに伴い，異なる言語間で文対応の付いた対訳コーパスの取扱いの重要性が認識されつつある．これまでの対訳コーパスからの専門用語等の抽出には，同時出現頻度といった統計情報を基に対応付けるアプローチがとられていたが，新たな語のような出現頻度が低い場合には必ずしも有効でない問題がある．本稿では，対訳コーパスに言語解析処理を適用し，文構造情報を用いた対訳辞書作成の方法を提案する．提案方法では，言語特性に基づいて単語を並べ替え，語順を揃えた対訳文間で対応付いた単語対を抽出する．日英対訳コーパスを用いた実験結果から，低頻度の単語についても対訳辞書作成が可能となることを示す．また，英日機械翻訳における単語アライメント学習の強化前後で比較した実験結果を通して提案方法による対訳辞書の効果を明らかにする．

Construction of Parallel Word Pair from Parallel Corpus Using Sentence Structure Information

Shogo SHINKAI[1,a)]　Daisuke SATO[1,b)]　Tsutomu MATSUNAGA[1,c)]

Abstract: With an effective utilization of information asset, an importance of handling existing limited corpus has been recognized. Conventional methods to extract parallel word pairs from parallel corpus, which have adopted statistical measures, have been indicated that they have a limitation in obtaining low frequency pairs of words such as new words. Using sentence structure information, a novel method to construct parrallel word pairs is proposed in this paper. Experimental results using Japanese-English parallel corpus demonstrate the effectiveness of the proposed method.

1. はじめに

情報資産の有効活用の動きに伴い，異なる言語間で文対応の付いた対訳コーパスの取扱いの重要性が認識されつつある [1]．一般に対訳データの収集には多大なコストと時間を要する [2] ことから，現有する限られたコーパスの効果的活用が技術課題となっている．

これまでの対訳コーパスからの専門用語等の辞書作成には，候補となる単語をあらかじめ準備しておき，それらを統計情報を基に対応付けるアプローチがとられている [3]．対応付けの統計情報には Dice 係数が挙げられ，対訳文間の同時出現頻度を基に対応関係の強い組を抽出する検討例が知られる．しかしながら，このアプローチによれば，対象の単語をあらかじめ準備する必要がある他，新たな語のような出現頻度が低い場合には必ずしも有効でない問題がある [4]．

対訳コーパスから翻訳機能を実現する統計的機械翻訳方式が実用されつつある [5], [6]．比較的定型の表現様式の下で確実な用語の訳が要求される技術マニュアル等のドメイン (自動車や IT 分野に代表的) の産業翻訳 [7] には，対訳コーパスから文体や用語を学習する統計的機械翻訳方式が

[1] 株式会社ＮＴＴデータ 技術開発本部
Research and Development Headquarters, NTT DATA CORPORATION
Toyosu Center Bldg. Annex, 3-9 Toyosu 3-chome, Koto-ku, Tokyo 135-8671, Japan
[a)] shinkais@nttdata.co.jp
[b)] satoudic@nttdata.co.jp
[c)] matsunagat@nttdata.co.jp

適合するといえる．ここに統計的機械翻訳方式の学習において，対訳コーパスにおける単語アライメント学習が訳質に関与し，出現頻度の低い単語は誤り易い傾向にあることが指摘されている [8]．

本稿では，対訳コーパスに言語解析処理を適用し，文構造情報を用いた対訳辞書作成の方法を提案する．提案方法では，言語特性に基づいて単語を並べ替え，語順を揃えた対訳文間で対応付いた単語対を抽出する．日英対訳コーパスを用いた実験結果から，低頻度の単語についても対訳辞書作成が可能となることを示す．また，英日機械翻訳における単語アライメント学習の強化前後で比較した実験結果を通して提案方法による対訳辞書の効果を明らかにする．本稿のねらいは，従来方法では対処し得なかった複数の単語の組が対訳文間で存在する場合に，提案方法により文構造情報を用いることで単語対抽出が可能になることを示す点にある．

2. 言語特性に基づく語順並べ替えと対訳用語辞書作成

これまで対訳コーパスからの専門用語等の単語対抽出には，対訳文間における同時出現頻度といった統計情報を基に対応付けるアプローチがとられ，その統計情報には Dice 係数が挙げられる [3]．Dice 係数は次式：

$$D(X,Y) = \frac{2 \cdot f_{XY}}{f_X + f_Y}$$

で定義され，単語間の対応関係が算出される．ここに，f_X および f_Y は，単語 X および単語 Y が独立に出現する頻度，f_{XY} は単語 X，Y が対訳文間で同時に出現する頻度である．

彼|は|、|東京|に|ある|会社|で|働い|て|いる|。
he|works|for|a|company|in|Tokyo|.

図 1a　対訳文の例
Fig. 1a　An parallel sectence example.

彼|は|、|東京|に|ある|会社|で|働い|て|いる|。
he|*ga*|Tokyo|in|company|for|works|.

図 1b　主辞後置英語による対訳文の例
Fig. 1b　An parallel sectence example by Head-Final English.

図 1a は対訳文の例である．日本語は ‘彼は、東京にある会社で働いている。’，英語は ‘He works for a company in Tokyo.’ で，図の ‘|’ は単語の区切りを表している．日本語の単語 ‘東京’，‘会社’，英語の単語 ‘Tokyo’，‘company’ の間では同時出現頻度 1 であり（**表 1**），このように対訳文間で複数の単語の組が存在する場合には区別がつかない．

表 1　単語対抽出の算出例
Table 1　A calculation example of word pair extraction.

日本語単語	英語単語	同時出現頻度	提案方法のスコア
東京	Tokyo	1	0.04
東京	company	1	0.29
会社	Tokyo	1	0.21
会社	company	1	0.04

提案方法では，日本語における主辞後置性（係り受け関係の係り先が後ろに位置する）という言語特性により，英語の主辞をあらかじめ句の末尾に並べ替えた主辞後置英語（Head-Final English:HFE）[9], [10] を導入する[*1]．**図 1b** に示されているように，日本語文と照らし合わせて語順が揃えられる[*2] ことがわかる．提案方法では語順が揃えられた対訳文（図 1b）で，単語の対応関係の強さを各単語の文中における相対的な出現位置の差分から算出する．この出現位置差分が小さいほど，対応関係が強い単語対として扱われる．

提案方法の出現位置差分のスコア A_{diff} は次式：

$$A_{diff} = \left| \frac{O_j}{N_j} - \frac{O_e}{N_e} \right|$$

で算出される．ここに N_j および N_e はそれぞれ，単語に分割された日本語および主辞後置英語の構成単語数，O_j および O_e はそれぞれ，注目する日本語単語および英語単語の文中における出現順である．図 1b の対訳文の例では，日本語の構成単語数 N_j=12，英語の構成単語数 N_e=8，‘東京’ の出現順 4，‘会社’ の出現順 7，‘Tokyo’ の出現順 3，‘company’ の出現順 5 であり，例えば，‘東京’ と ‘Tokyo’，については，$\left|\frac{4}{12} - \frac{3}{8}\right| = 0.04$ が得られる．‘東京’，‘会社’，‘Tokyo’，‘company’ の語における単語対に対して算出されたスコアを表 1 に示す．表に示されているように，‘東京’，と ‘Tokyo’，‘会社’ と ‘company’ が文における語の位置が考慮されることを通して他の組に比べて相対的に小さい値となり，対応関係の強さがスコアに反映されていることがわかる．対訳辞書は，あらかじめ定められたしきい値を通して選定された単語対から作成される．

これまで述べてきたように，提案方法では言語特性から単語を並べ替えて変換した中間的単語列を扱い，対訳文間で語順が整合することを基に，出現位置から対応付いた単語対抽出により対訳辞書作成が行われる．

[*1] 主辞後置英語（HFE）は係り受け解析 [11] を基に，主辞（係り先）の後ろへの移動とあわせて，助詞（‘ga’，‘o’）の補完，冠詞の削除を通して得られる．
[*2] 片方の単語列を言語学的特徴を基に単語を並べ替えて変換した中間的単語列を扱う [12] ことで語の対応付けを図る手法が知られる．日英対訳コーパスに対して日本語の語順との整合性が考慮された主辞後置英語を扱うことにより，対訳文の質の定量化が検討されている [1]．

3. 実験

3.1 文構造情報を用いた対訳コーパスからの単語対抽出

本節では英語学習で扱われる基本英単語の用例文 [13] からの日英対訳コーパス 2,613 文対（**表 2**）を用いる[*3]．同時出現頻度の上位 25 の単語対を**表 3a** に示す[*4]．また，**表 3b** は同時出現頻度 1 の例であり，Dice 係数の小さいものから 10 件を示した．表からわかるように同時出現頻度が高い単語対はいずれも正しく（表 3a），同時出現頻度 1 と低くなると，例えば表 3b の 2 番目 ‘判断’ と ‘bad’ のように誤った単語対が含まれるようになることがわかる．

表 2 実験データ（日英対訳コーパス 2,613 文対）
Table 2 Japanese-English parallel corpus (2,613 parallel sentences).

言語	日本語	英語
異なり語数	4,423	3,309
全語数	35,513	25,327

同時出現頻度 1 で 1 対 1 の関係にない単語対は 1,474 件で，提案方法の処理対象とした．**表 4** に単語の組の数の分布が示されているように，最大 5 つの重複[*5] がみられている．

表 4 単語の組の数の分布
Table 4 Distribution of multiple word pairs.

単語の組の数	件数
2	467
3	133
4	32
5	3

表 5 は抽出単語対の例である．しきい値 0.15 とすることにより 1 対 1 の単語対として 378 件を抽出した．表中の ‘×’ のマークで示されているように誤りが含まれており，目視により 51.9%（378 件中に正しいのは 196 件）の精度である．表の 3 番目に ‘望む’ と ‘anything’ の誤った単語対がみられるが，これは**表 6** の 1 つ目の対訳文において，英語の ‘anything+主語+動詞’ の表現に対して日本語の ‘なんでも’ が後に位置するためで，主語の欠落による不整合もみられる．また，誤った単語対取得の典型例として次が挙げられる．表 6 の 2 つ目の対訳文で，‘容器’ と ‘juice’ の組（スコア 0.08）が選定されているが，これは動詞にかかる修飾語句の出現順において日本語の文と整合しないことに起因している．日本語の文として，‘多くのジュースがガラス瓶よりプラスティック容器で、売られています。’ が望ましいものであるといえる．この統語構造への対応は検討課題に挙げられる．

以上から，同時出現頻度といった従来方法では抽出し得なかった単語対が，提案方法により文構造情報を用いることで抽出可能となるが，5 割強の精度であることが示された．

3.2 抽出単語対を用いた対訳アライメント学習強化

本節では英語から日本語への機械翻訳（英日機械翻訳）を取り上げ，提案方法により抽出された単語対の訳質に与える効果を評価する．元の対訳コーパスに単語対を付け足すこと (対訳文増加の形で) を通して単語アライメント学習[*6] を強化し，学習強化前後を比較する[*7]．本実験ではソフトウェアマニュアル（プログラミング言語）の Python 対訳データ [16] を用いる．

評価のためのテストセットに 1,000 文対，開発セットに 1,000 文対を選抜し，学習に用いる訓練セットは 24,215 文対である．単語対抽出には訓練セットのうちから文字数が少なく章や節に該当するものを除いた 7,297 文対を用いた．前節と同じく，同時出現頻度 1 で 1 対 1 の関係にない単語対から提案方法によりしきい値 0.15 を通して単語対 1,289 件を得た．提案方法の学習強化ではこの 1,289 単語対が追加された 25,504 文対相当の対訳コーパスを扱うこととなる．なお，1 対 1 の関係で得られる単語対は 2,088 件（同時出現頻度最大 1513，最小 1）あり，これを従来方法の学習強化での追加単語対に扱った．

表 7a 翻訳の自動評価結果
Table 7a Translation evaluation results by automatic metrics.

	BLEU(%)	RIBES(%)	未知語件数
学習強化前	32.5	74.4	563
学習強化後（従来方法）	32.3	74.2	550
学習強化後（提案方法）	32.5	74.1	553

表 7b RIBES 値変動に関する文数
Table 7b Number of sentence regarding RIBES score changes.

	RIBES 値向上	RIBES 値不変	RIBES 値低下
従来方法	267	463	270
提案方法	331	356	313

[*3] 言語解析処理については文献 [5] の方式を採用している．
[*4] 単語対の取得には，注目する日本語の単語との Dice 係数が最大となる英語の単語を取り出した上で，その英語の単語との Dice 係数最大がその注目した日本語の単語と一致した際に取得する方法 [3] を採用している．
[*5] 例えば ‘合計’ に対して ‘total’, ‘1.6’, ‘trillion’, ‘tax-reduction’, ‘proposes’ との 5 つの組が得られている．
[*6] 単語アライメント学習は，対訳コーパスにおける言語間で単語が翻訳される期待値最大化の処理により実現される．本稿の実験では単語アライメントツールに GIZA++ [14] を用いている．
[*7] 統計的機械翻訳の Moses ツールキット [15] を使用し，各モデルの重みは開発セットを用いた誤り最小化学習 (MERT) により最適化した．デコーダには文献 [6] のものを使用した．

テストセットにおける単語アライメント学習強化前後の訳質の自動評価による結果を**表 7a** に示す．翻訳結果に含まれる未知語件数もあわせて示した．訳質評価は機械翻訳の自動評価指標としてよく知られる BLEU [17] および RIBES [18] [*8] を採用した．これら指標は大きな値ほど高い訳質を指す．

表に示されている通り，学習強化により未知語件数の減少はみられているが，自動評価指標値によれば従来方法と提案方法のいずれも訳質の向上はみられていない．これは追加単語対に含まれる誤りが影響しているものと推察される．従来方法と提案方法を比べると，BLEU 値では提案方法の方が上回る一方，RIBES 値では逆に提案方法の方が下回る結果が示されている．ここにテスト文 1,000 件において，文毎の値が得られる RIBES 値により，学習強化前後の値の変動を調査した（**表 7b**）．従来方法では値の低下する文数が向上に比べて多いのに対し，提案方法では向上の文数が低下を上回る結果が得られている．追加単語対の数が従来方法で 2,088 件（同時出現頻度が高いものが多く含まれる），提案方法で 1,289 件（同時出現頻度 1 の重複する組に限定）であり，低頻度の単語へ学習強化する提案方法の効果を相対的にみることができる．

表 8 は，提案方法の学習強化により訳質に向上がみられた翻訳文の例である．翻訳対象文の単語に追加単語対の該当はないことが特筆すべき点であり，低頻度の単語対追加が間接的に構文に関わる訳質向上へも寄与することを示唆している．

上述したように訳質向上には単語対に含まれる誤りの除去が必要であり，目視による選定といった対処が考えられる．また前節で挙げたように，対訳文における統語構造 [19] が考慮された対応が検討課題に挙げられる．

4. おわりに

本稿では，対訳コーパスに言語解析処理を適用し，文構造情報を用いた対訳辞書作成の方法を提案した．提案方法では，言語特性に基づいて単語を並べ替え，語順を揃えた対訳文間で対応付いた単語対を抽出することで，低頻度の単語についても対訳辞書作成が可能となることを示した．英日機械翻訳の実験結果を基に，単語アライメント学習の強化前後の比較を通して提案方法による対訳辞書の効果を明らかにした．

統計的機械翻訳適用において大量の対訳コーパスを要することが課題となっているが，現有する限られたコーパスの効果的活用により適用ドメイン拡大につながる技術開発について引き続き取り組む予定である．

[*8] BLEU は正解訳との N グラム (N=1,2,3,4) の適合率の幾何平均で，自動評価尺度のデファクトスタンダードに該当するものである．RIBES は語順が重視された指標で，日英等のように語順が大きく異なる言語間での訳質評価に有効とされるものである．

参考文献

[1] 松永務，新海正吾，末永高志：翻訳メモリのクリーンアップのための対訳文ランキング，知能と情報, Vol. 27, pp. 621–625 (2015).
[2] 石坂達也，内山将夫，隅田英一郎，山本和英：大規模オープンソース日英対訳コーパスの構築，2009-NL-191，pp. 1–6 (2009).
[3] 北村美穂子，松本裕治：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, pp. 727–736 (1997).
[4] 山田節夫，中岩浩巳，池原悟：対訳コーパスから対応する表現対の自動抽出，言語処理学会第 2 回年次大会発表論文集，pp. 197–200 (1996).
[5] 須藤克仁，鈴木潤，秋葉康弘，塚本元，永田昌明：英中韓から日本語への特許文向け統計翻訳，言語処理学会第 20 回年次大会発表論文集，P6-8, pp. 606–609 (2014).
[6] 秋葉泰弘, 我妻光洋, 荒井和博：AAMT-MT フェア 2014 展示報告-特許翻訳などの専門的な外国語文書も自然な日本語に-多言語統計翻訳プラットフォーム-，AAMT Journal，57, pp. 70–71 (2014).
[7] 西野竜太郎：ソフトウェア・ローカリゼーションのこれから，日本翻訳ジャーナル，262, pp. 60–64 (2012).
[8] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Goldsmith, M. J., Hajic, J., Mercer, R. L. and Mohanty, S.: But dictionaries are data too, *Proc HLT'93*, pp. 202–205 (1993).
[9] 磯崎秀樹：英日翻訳における語順について，言語処理学会第 16 回年次大会発表論文集，B4-2, pp. 884–887 (2010).
[10] Isozaki, H., Sudoh, K., Tsukada, H. and Nagata, M.: Head finalization: A simple reordering rule for SOV languages, *Proc WMT-MetricsMATR*, pp. 244–251 (2010).
[11] Suzuki, J., Isozaki, H., Carreras, X. and Collins, M.: An emprical study of semi-supervised structured conditional models for dependency parsing, *Proc ACL-EMNLP*, pp. 551–560 (2009).
[12] Collins, M., Koehn, P. and Kucerova, I.: Clause restructuring for statistical machine translation, *Proc ACL'05*, pp. 531–540 (2005).
[13] VOA Special English Word Book http://www.manythings.org/voa/words.htm
[14] GIZA++ http://www.fjoch.com/GIZA++.html
[15] Moses http://www.statmt.org/moses/
[16] Python 対訳データ http://www2.nict.go.jp/univ-com/multi_trans/member/mutiyama/manual/index-ja.html
[17] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU:A method of automatic evaluation of machine translation, *Proc ACL*, pp. 311–318 (2002).
[18] Isozaki, H., Hirano, T., Duh, K., Sudoh, K. and Tsukada, H.: Automatic evaluation of translation quality for distant language pairs, *Proc EMNLP*, pp. 944–952 (2010).
[19] 中岩浩巳：対訳コーパス中の規則獲得不適文対の自動認定，情報処理学会第 57 回全国大会講演論文集，5R-08, pp. 269–270 (1998).

表 3a　単語対抽出の例
Table 3a　An example of parallel word extraction.

同時出現頻度	Dice 係数	日本単語	英単語
355	0.737	彼	he
275	0.764	彼女	she
153	0.576	私	I
105	0.582	あなた	you
73	0.789	彼ら	they
63	0.906	大統領	president
61	0.726	新しい	new
56	0.778	言っ	said
53	0.675	家	house
48	0.873	車	car
46	0.773	2	two
44	0.553	我々	our
42	0.604	多く	many
35	0.636	人々	people
31	0.602	年	years
31	0.939	合衆国	states
31	0.925	警察	police
30	0.870	水	water
29	0.951	犬	dog
28	0.836	本	book
26	0.881	政府	government
26	0.897	3	three
25	0.806	学校	school
25	0.893	会社	company
24	0.608	男	man

表 3b　単語対抽出の例（同時出現数 1）
Table 3b　An example of parallel word extraction(co-occurrence of one).

同時出現頻度	Dice 係数	日本単語	英単語
1	0.222	終結	end
1	0.250	判断	bad
1	0.250	手伝わ	let
1	0.250	まずい	bad
1	0.286	立ち止まっ	field
1	0.286	野原	field
1	0.286	望ん	proposals
1	0.286	敵地	position
1	0.286	大騒ぎ	wild
1	0.286	打ち方	hit

表 5　提案方法による単語対抽出の例
Table 5　An example of parallel word extraction by the proposed method.

スコア	日本単語	英単語	正誤
0.29	薄く	thin	○
0.29	濃く	thick	○
0.40	望む	anything	×
0.40	生き残っ	even	×
0.40	社員	employee	○
0.50	不誠実	dishonest	○
0.50	不正直	dishonest	○
0.50	調べ	relationship	×
0.50	聴き	listen	○
0.50	贈り物	pleased	×
0.50	選ぶ	choose	○
0.50	先住民	native	○
0.50	生き残ら	crashed	×
0.50	込み入っ	complex	○
0.50	合意	settlement	○
0.50	割る	split	○
0.50	解か	removed	×
0.50	温まる	once	×
0.50	リーダー	opponent	×
0.50	どなっ	shouted	○

表 6　実験で用いた対訳文の例
Table 6　An example of parallel sentence used in the experiment.

日本語文	あなたが望むことをなんでも話し合いましょう。
英語文	We can discuss anything you wish.
英語文（HFE）	we *ga* anything you wish *o* discuss can.
日本語文	ガラス瓶よりプラスティック容器で、多くのジュースが売られています。
英語文	More juice is sold in plastic containers than in glass bottles.
英語文（HFE）	more juice *ga* glass bottles in than plastic containers in sold is.

表 8　英日機械翻訳例
Table 8　A translation example.

原文	Arguments are converted to those units:
参照訳	引数は以下のようにして変換されます:
翻訳結果（学習強化前）	引数：これらの単位に変換されます。
翻訳結果（学習強化後）	引数は以下のもの単位に変換されます。