2-3　パラレルコーパスの自動生成技術

2-3 Automatic Construction Technology for Parallel Corpora

内山将夫　谷村　緑
UTIYAMA Masao and TANIMURA Midori

要旨

大規模な日英対訳コーパスを作ることを目的として、1989 年から 2001 年までの読売新聞と The Daily Yomiuri とから日英記事対応と文対応とを得た。そのときの方法は、まず、内容が対応する日本語記事と英語記事とを言語横断検索により得て、次に、その対応付けられた日英記事中にある日本語文と英語文とを DP マッチングにより対応付けるというものである。しかし、それにより対応付けられた記事対応や文対応には、間違った対応（ノイズ）が多く含まれる。そのため、我々は、本稿において、そのようなノイズを避けて、正しい対応のみを得るための信頼性の高い尺度を提案し、その信頼性の評価をした。実験の結果、我々の提案した尺度を用いることにより、良質な記事対応や文対応が得られることが分かった。

We have aligned Japanese and English news articles and sentences, extracted from the Yomiuri and the Daily Yomiuri newspapers, to make a large parallel corpus. We first used a method based on cross-lingual information retrieval to align the Japanese and English articles and then used a method based on dynamic programming (DP) matching to align the Japanese and English sentences in these articles. However, the articles and sentences included many incorrect alignments. To remove these, we propose two measures that evaluate the validity of the alignments. Using these measures, we successfully extracted valid article and sentence alignments.

[キーワード]
日英対訳コーパス，記事アライメント，文アライメント
Japanese-English parallel corpus, Article alignment, Sentence alignment

1　まえがき

日英対訳コーパスは、機械翻訳などの自然言語処理において必要であるばかりでなく、英語学や比較言語学、あるいは英語教育や日本語教育などにとっても非常に有用な言語資源である。しかしながら、これまで、一般に利用可能で、かつ、大規模な日英対訳コーパスは存在していなかった。

そのような背景の中で、我々は、比較的大規模な日本語新聞記事集合及びそれと内容的に一部対応している英語新聞記事集合とから、大規模な日英対訳コーパスを作ることを試みた。

そのための方法は、まず、内容が対応する日本語記事と英語記事とを得て、次に、その対応付けられた日英記事中にある日本語文と英語文とを対応付けるというものである。

ここで、我々が対象とする日本語記事と英語記事においては、英語記事の内容が日本語記事の内容に対応している場合には、その英語記事は、日本語記事を元にして書かれている場合が多いのであるが、その場合であっても、日本語記事を直訳しているわけではなく、意訳が含まれていることが多く、さらに、日本語記事の内容の一部が英語記事においては欠落していたり、日本語記事にない内容が英語記事に書かれている場合もある。また、記事対応付けを得るための日本語記事集合と英語記事集合についても、英語記事集合の大きさは日本語記事集合の大きさの 6 %未満であるので、日本語記事の中で、対応する英語記事があるものは極少数である。

そのため、記事対応付け及び文対応付けにあたっては、非常にノイズが多い状況の中から、適切な対応付けのみを抽出しなくてはならないので、対応の良さを判断するための尺度は信頼性の高いものでなくてはならない。

本稿では、そのような信頼性の高い尺度を、記事対応付けと文対応付けの双方について提案し、その信頼性の程度を評価する。

以下では、まず、対応付けに用いた日英新聞記事について概要を述べ、次に、記事対応付けの方法と文対応付けの方法を述べたあとで、それぞれの対応付けの精度を評価する。

2 対応付けに用いた日英新聞記事

対応付けの元データは、日本語記事は「読売新聞」、英語記事は「The Daily Yomiuri」であり、それぞれ 1989 年 9 月から 2001 年 12 月までの記事を利用した。この期間における総記事数は、日本語記事は約 200 万であり、英語記事は約 11 万である。このように、英語記事の方が少ないので、対応付けにおいては、各英語記事に対応する日本語記事を求めることにした。

記事のメタ情報として、The Daily Yomiuri には、1996 年 7 月中旬から、「本紙翻訳＝Y/N」という情報が各記事に付いている。これは、その英語記事を書くにあたって、読売新聞の記事を元にしたかどうかという意味であるので、1996 年 7 月中旬からは、「本紙翻訳＝Y」である英語記事についてのみ、対応する日本語記事を求めることにした。このときの英語記事の数は 35318 である。一方、1996 年 7 月中旬以前には、そのような情報はないので、すべての英語記事について対応する日本語記事を求めることにした。このときの英語記事の数は 59086 である。なお、以下では、1996 年 7 月中旬以前の記事集合を「1989-1996」と書き、1996 年 7 月中旬以降の記事集合を「1996-2001」と書く。

1989-1996 については、全英語記事を利用するため、1996-2001 と違って、そもそも、各英語記事について対応する日本語記事がない場合がある。そのため、どのくらいの英語記事に、対応する日本語記事があるかを推測するために、「本紙翻訳＝Y」の割合を、1997 年から 2001 年の記事について調べたところ、67.9 %であった。

対応を求めるにあたって、各英語記事に対応する日本語記事は、互いに近い日付であると考えられる。そのため、各英語記事について、その日付の前後 2 日の範囲の日本語記事の中から対応する記事を見付けることにした。このとき、1 日分の英語記事について、日本語記事は 5 日分があるが、このときの平均記事数は、1989-1996 については、英語記事が 24、日本語記事が 1532、1996-2001 については、英語記事が 18、日本語記事が 2885 である。

このように、非常に曖昧性があり、かつ、対応記事も場合によっては存在しないという、ノイズの多い状況の中から対応記事を見つける必要があるので、信頼性の高い記事対応（評価）尺度が必要である。また、文対応についていえば、たとえ記事同士が対応していたとしても、その対応は、直訳関係にあるものは少なく、どちらかというと、日本語記事を材料として英語記事を書いたというような状況である。そのため、直訳に近い文対応を抽出するためには、信頼性の高い文対応（評価）尺度が必要である。

3 ベースラインとなる記事対応付け及び文対応付けの方法

記事対応付けは、言語横断検索の枠組みで行う。つまり、英語記事を質問とし、それに関連する記事を日本語記事データベースから検索することにより、与えられた英語記事と対応する日本語記事を見付ける。

このとき、一般に、質問である英語記事を日本語に変換するか、あるいは、データベースである日本語記事を英語に変換する必要がある。本研究では、データベースである日本語記事を英語（の単語集合）に変換した。すなわち、まず、日本語記事を茶筌により形態素解析し、形態素解析された結果の単語を EDR 辞書等を利用して英語に変換した。

いったん、日本語記事が英単語集合に変換されてしまえば、あとは、通常の情報検索と同様にして、質問として与えられた英語記事に最も類似するような日本語記事（の英単語集合への変換結果）を検索することができる。そして、その日本語記

事をもって対応記事とする。このときの英語記事と日本語記事の類似度としては、情報検索に有用な尺度として知られている BM25[1] を利用した。

BM25 により対応付けられた日英記事における文間の対応は DP マッチングで求めた[2][3]。DP マッチングで文対応を得るアルゴリズムの簡潔な記述は文献[3] を参照のこと。ここでは、日本語文（集合）から得られた内容語集合 J と英語文（集合）から得られた内容語集合 E との類似度、SIM(J, E) についてのみ述べる。類似度 SIM は以下のように定義される。

$$SIM(J, E) = (co(J \cap E) + 1) / (|J| + |E| - 2\,co(J \cap E) + 2)$$

ただし、|J| と |E| は日本語文集合 J と英語文集合 E に含まれる単語の数である。また、co(J∩E) は、J 中の単語と E 中の単語とで 1 対 1 対応が付いた単語の数である。ただし、日英の単語の一対一対応を求めるためには、EDR 日英辞書及び EDR 英日辞書を利用した。

以上のように定義された類似度 SIM を用いて、文対応を付けたが、このとき、文対応付けに用いたプログラムでは、DP マッチングにおける文間の対応としては、1 対 n もしくは n 対 1、ただし、$1 \leqq n \leqq 6$ しか許していない。この条件下で、文対応プログラムの精度を、人手により文対応が付けられている白書データに適用することにより求めたところ、98 %以上であった。すなわち、白書データのように、日本語が忠実に英語に訳されているようなデータについては、文対応プログラムの精度は十分に高いといえる。

4　信頼性の高い記事対応尺度と文対応尺度の提案

3 において、記事対応の類似度 BM25 と文対応の類似度 SIM とを導入した。しかしながら、これらの類似度のみを利用して記事対応や文対応を付けた場合には、以下の実験で示すように、十分に精度の高い記事対応や文対応を得ることはできない。そのため、本節では、記事対応と文対応の双方について、信頼性の高い、新たな尺度を定義する。

まず、記事対応についてであるが、我々は、日本語記事 J と英語記事 E の類似度として BM25 (J, E) を導入した。この類似度は、単語集合間の類似度であるので、文の順序などは考慮できない。そのため、文の順序を考慮できる記事対応尺度として、AVSIM (J, E) を定義する。これは、J と E との文対応を {(J1, E1), … , (Jm, Em)} としたとき、以下の式である。

$$AVSIM(J, E) = (SIM(J1, E1) + \ldots + SIM(Jm, Em)) / m$$

AVSIM が高い値となるのは、個々の文対応の類似度 SIM が高い場合であるので、そのような場合には、記事としての対応も良いと考えた。

次に文対応の良さの尺度について述べる。**3** で述べたように、我々の文対応付けプログラムの精度は、白書データのように日本語文と英語文とが原文と訳文という関係にあるようなものを対応付ける限りにおいては、高精度である。しかし、**2** で述べたように、日本語記事と英語記事との関係は、一般には、原文と訳文という関係ではない。そのため、**3** の方法で文対応付けをした場合には、適切な対応とともに不適切な対応も多く得られる。そのようにノイズの多い状況から、適切な対応のみを抽出するためには、文対応の尺度として、文類似度だけでなく、記事対応の尺度も利用すれば良いと考えた。そのため、日本語記事 J と英語記事 E との記事対応における、文 Jk と Ek との文対応尺度として、

$$SntScore(Jk, Ek) = AVSIM(J, E) \times SIM(Jk, Ek)$$

を定義した。この尺度は、同一記事対応内で文対応を比べる場合には文類似度 SIM と同じ順位を与えるが、異なる記事間での文対応の比較では、文類似度だけでなく、記事対応の尺度値も高いような文対応を優先する。

5　記事対応付けの精度

5.1　無作為抽出による精度評価

記事対応付けは、各英語記事との類似度 BM25 が高い日本語記事を検索することによりなされる。このとき、類似度 1 位の日本語記事についての記事対応付けの精度を 1996-2001と 1989-1996 とについて表 1 に示す。

表1 類似度 1 位の記事対応の精度

評価値	1996-2001			1989-1996		
	下限	割合	上限	下限	割合	上限
A	0.49	0.59	0.69	0.20	0.29	0.38
B	0.06	0.12	0.18	0.08	0.15	0.22
C	0.03	0.08	0.13	0.03	0.08	0.13
D	0.13	0.21	0.29	0.38	0.48	0.58

表 1 において、「評価値」とは、記事対応の良さの人手による判定の評価値であり、その基準は、A は「記事全体の記述の 5〜6 割程度以上について意味の対応がとれる」、B は「2〜3 割程度以上 5〜6 割程度以下について意味の対応がとれる」、D は「全然違う」、C は「A，B，D 以外」である。「割合」とは、1996-2001 と 1989-1996 のそれぞれから、100 記事対応ずつを一様無作為抽出したときに、その評価値であった記事対応の割合である。「下限」「上限」とは、割合の 95 %信頼区間の下限と上限である。

2 で述べたように、1996-2001 については、「本紙翻訳＝Y」なる英語記事のみを対象としたが、1989-1996 については、全英語記事を対象とした。そのため、1989-1996 の精度は、1996-2001 よりも低い。また、1996-2001 の精度が 1989-1996 の精度よりも高いといっても、それでも、評価値 A が約 60 %、A もしくは B が約 70 %であるので、BM25 による記事対応付けの結果をそのまま利用した場合には、ノイズとなる記事対応が多すぎる。

我々の観察によれば、評価値が A もしくは B の記事対応は、そこから日英言語表現間の対応が抽出できそうという意味において、有用な記事対応である。このような記事対応のみを抽出するには、BM25 による記事対応付けの結果をそのまますべて利用するのではなく、対応の良さにより対応付けの結果をソートし、その上位のみを抽出すれば良い。

5.2 ソートした場合の記事対応の精度

記事対応の良さの指標として、AVSIM と BM25 のどちらが適当かを比較した。表 1 と同じデータに対して、それぞれの値の降順により記事対応をソートし、評価値が A もしくは B の場合を正解とし、各順位までにおける正解の個数とその割合とを調べた。それを表 2 に示す。表 2 から、我々は、AVSIM の方が BM25 よりも、記事対応の良さとして適切な尺度であると判断した。

表2 順位と精度

順位	1996-2001				1989-1996			
	AVSIM		BM25		AVSIM		BM25	
	数	割合	数	割合	数	割合	数	割合
5	5	1.00	5	1.00	5	1.00	2	0.40
10	10	1.00	8	0.80	10	1.00	4	0.40
20	20	1.00	16	0.80	19	0.95	9	0.45
30	30	1.00	25	0.83	28	0.93	16	0.53
40	40	1.00	34	0.85	34	0.85	24	0.60
50	50	1.00	39	0.78	37	0.74	28	0.56
60	60	1.00	47	0.78	42	0.70	30	0.50
70	66	0.94	55	0.79	42	0.60	35	0.50
80	70	0.88	62	0.78	43	0.54	38	0.47
90	71	0.79	68	0.76	43	0.48	40	0.44
100	71	0.71	71	0.71	44	0.44	44	0.44

AVSIM の精度の方が BM25 の精度よりも高い理由は、**4** で述べたように、AVSIM が、BM25 と違って、個々の文対応の良さまでも考慮した尺度であるからと考える。AVSIM を利用することにより、ノイズの多い記事対応の中から、良質な記事対応のみを抽出することが可能となる。

6 文対応付けの精度

2 で述べたように、たとえ、日英記事間に内容上の対応があったとしても、文間対応があるとは限らないので、対応付けられた記事から得られる文対応はノイズが多いものとなる。そのため、BM25 による類似度 1 位の記事対応すべてから得られる文対応すべてを SntScore により降順にソートし、その上位のみを利用することにより対応の良いものを抽出することにした。

このような文対応の数は、1989-1996 と 1996-2001 を合わせた全体で、約 130 万だけある。文対応の中では、1 対 1 対応が最も重要である。また、文対応といっても、新聞記事には、中見出しなどの、必ずしも文でないものもある。そのため、1 対 1 対応の中で、文末が句点やピリオドなどで終っているもののみを取り出し、これを特に「1：1」と呼び、その他の対応を「1：n」と呼ぶことにする。1：1 の数は、約 64 万ある。1：n の数は、

表3　順位と1：1の精度

範囲	o数	x数
1 -	100	0
30001 -	99	1
60001 -	99	1
90001 -	97	3
120001 -	96	4
150001 -	92	8
180001 -	82	18
210001 -	74	26
240001 -	47	53
270001 -	30	70

約66万ある。

1：1の精度を求めるために、SntScoreにより降順にソートされた上位30万対応について、3万対応ごとに100ずつを一様無作為抽出した。この各対応について、x/oの2値評価をした。ここで、xは「意味が全然違う」であり、oは「意味が全然違うことはない」である。その結果のx/oの数を表3に示す。

表から分かるように、順位が下っていくにつれて、xの数が指数的に増加している。このことは、SntScoreが、効率良く、適切な1：1を上位に順位付けていることを示している。表3から、15万対までは十分に信頼できる対応であると言える。なお、15万対までのoの累積の割合は0.982である。

次に、1：nの精度を求めるために、SntScoreにより降順にソートされた上位について、表3の「1-90000」「90001-180000」「180001-270000」の各範囲について、それらの1：1のSntScoreの範囲に収まるような1：nの精度を求めた。精度を求めるときには、1：1のときと同様に、各範囲から100対を一様無作為抽出し、x/oの2値評価をした。その結果を表4に示す。表より、「1-90000」の範囲の38090個の1：nについては、精度の良い対応であると言える。

以上述べたように、SntScoreにより文対応をソートすることにより、1：1と1：nの双方について、上位には、十分に精度の高い文対応が得られる。なお、SntScoreの精度の方がSIMの精度よりも高いことも確認している。SntScoreの精度の方が高い理由は、**4**で述べたように、SntScoreが、SIMと違って、記事対応の良さまでも考慮した尺度であるからと考える。

表4　順位と1：nの精度

範囲	1:nの数	o数	x数
1 -	38090	98	2
90001 -	59228	87	13
180001 -	71711	61	39

7　データ公開

我々は、**6**で述べた文対応について、1：1の上位15万対と1：nの上位3万対とを、読売新聞社からの許可を得て、2002年より教育及び研究目的に公開しており、現在までに、100を超える機関や個人からデータ入手の申込みを受けた。このデータは、機械翻訳や英語教育[4]等に利用されている。また、我々は、このデータを検索できるサイトとして「言の場」(http://www.kotonoba.net/~snj/cgi-bin/text-search/text-search.cgi)を開設している。

8　むすび

ノイズの多い日英新聞記事集合から、内容が対応した記事対応と文対応を得るための信頼性の高い尺度を提案した。それら尺度を用いることにより、1989年から2001年までの読売新聞とThe Daily Yomiuriとから記事対応と文対応を得た。それらの中で、比較的良質と推定された文対応は、1対1対応が約15万あり、1対1対応以外が約3万8千ある。これらは、一般に公開され、教育研究目的に役立っている。

参考文献

1 S. E. Robertson and S. Walker, "Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval", SIGIR, pp.232-241, 1994.

2 William A. Gale and Kenneth W. Church, "A Program for Aligning Sentences in Bilingual Corpora", Computational Linguistics, 19:1, pp.75-102, 1993.

3 Takehito Utsuro, Hiroshi Ikeda, Masaya Yamane, Yuji Matsumoto, and Makoto Nagao, "Bilingual Text Matching using Bilingual Dictionary and Statistics", COLING, pp.1076-1082, 1994.

4 Kiyomi Chujo, Masao Utiyama, and Shinji Miura, "Using a Japanese-English Parallel Corpus for Teaching English Vocabulary to Beginning-Level Students", English Corpus Studies, 13, 153-172, 2006.

内山将夫（うちやまさお）
知識創成コミュニケーション研究センター自然言語グループ主任研究員（旧情報通信部門けいはんな情報通信融合研究センター自然言語グループ主任研究員） 博士（工学）
自然言語処理

谷村 緑（たにむらみどり）
京都外国語大学講師（元情報通信部門けいはんな情報通信融合研究センター自然言語グループ専攻研究員）
博士（言語文化学）
英語教育