情 報 学 基 礎　68－3
自 然 言 語 処 理 151－3
（２００２．９．１７）

日英新聞記事の対応付けと精度評価

内山将夫 井佐原均
通信総合研究所
{mutiyama,isahara}@crl.go.jp

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とから記事対応と文対応を得た．それらのなかで，比較的良質と推定されるものが，記事対応は約4万7千あり，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，現時点で一般に利用できる日英2言語コーパスとしては最大のものである．

Alignment of Japanese-English News Articles and Sentences

Masao Utiyama and Hitoshi Isahara
Communications Research Laboratory
{mutiyama,isahara}@crl.go.jp

We aligned Japanese and English news articles and sentences to make a large parallel corpus. The Japanese and English articles are extracted from the Yomiuri and the Daily Yomiuri, respectively. We showed that about 47 thousands article pairs, 150 thousands 1-to-1 sentence pairs, and 38 thousands 1-to-many sentence pairs are valid. Thus, we have built the largest Japanese-English parallel corpus, which is available to the public.

1 はじめに

大規模な日英対訳コーパスを作ることを目的として，互いに対応関係にあるような日英新聞記事を得た．そして，対応付けられた記事中にある日本語文と英語文とを対応付けることにより，文単位の対応を得た．

これらの対応のうち，比較的良質な対応と推定されるものが，記事対応については約4万7千，文対応については，1対1対応のものが約15万，1対1対応以外のものが約3万8千ある．これらは，現時点で一般に利用できる日英2言語コーパスとしては最大のものである．

本研究の貢献は，以下の3点である．(1) 記事対応および文対応のスコアとして信頼性の高いものを提案した．そのため，これらのスコアを使うことにより，今後，記事対応付けや文対応付けをする場合に，ノイズの多い対応のなかから信頼性の高い対応付けを抽出できるようになった． (2) 一般に利用可能な日英新聞記事を元データとしたため，そのデータを所有している人であれば，本研究の結果を利用することにより，日英の記事対応付けと文対応付けの結果を利用できるようになった[1]．(3) 対応付けにおいてコアとなるソフトウェアである文対応付けプログラムを一般に利用可能とした[2]ため，同様な研究をするのが容易になった．

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，それぞれの精度を述べる．最後に関連研究と結論を述べる．

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「読売新聞記事データ」における1989年9月から2001年12月までの記事を利用した．この期間における年間の記事数は，日本語記事は10万から35万程度であり，英語記事は4千から1万3千程度である．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，全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」と書くことにする．

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 (1996-2001についてもその可能性はある)．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なか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スコアが必要である．

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．たとえば，以下の例では，英語と日本語とで，「<★ s1>」により3対4に複雑に絡みあう対応があり，その間に「<◆ s2>」による対応がある．

> <★ s1> Two bullet holes were found at the home of Kengo Tanaka, 65, president of Bungei Shunju, in Akabane, Tokyo, by his wife Kimiko, 64, at around 9 a.m. Monday. </★ s1> <◆ s2> Police suspect right-wing activists, who have mounted criticism against articles about the Imperial family appearing in the Shukan Bunshun, the publisher's weekly magazine, were responsible for the shooting. </◆ s2> <★ s1> Police received an anonymous phone call shortly after 1 a.m. Monday by a caller who reported hearing gunfire near Tanaka's residence. </★ s1> <★ s1> Police found nothing after investigating the report, but later found a bullet in the Tanakas' bedroom, where they were sleeping at the time of the shooting. </★ s1>

[1]対応付けの結果を利用したい方は第1著者まで連絡して下さい．ただし，読売新聞と The Daily Yomiuri の CDROM は各自で用意する必要があります．

[2]http://www.crl.go.jp/jt/a132/members/mutiyama/softwares.html よりダウンロードできる．

<★ s1> 二十九日午前八時五十五分ごろ、東京都北区赤羽西四文芸春秋社長、田中健五さん（６５）方の二階東側外壁に、短銃で撃たれた跡があるのを、妻喜美子さん（６４）が見つけた。 </★ s1> <★ s1> 赤羽署で調べたところ、寝室の外壁に二か所の穴が確認され、銃弾一発が寝室内から発見された。 </★ s1> <★ s1> これに先立ち、午前一時すぎ、田中さん方周辺で「短銃の発射音のような音が二、三発聞こえた」という匿名の通報が同署にあり、署員が確認に向かったが、この時点で銃痕は発見できなかった。 </★ s1> <★ s1> 発射音がしたころ、田中夫妻は寝室で就寝中だったという。 </★ s1> <◆ s2> 同社が発行している週刊誌「週刊文春」が、最近、皇室批判記事を掲載していたことから、同署では、皇室批判に反発する右翼の犯行の可能性があるとみて、捜査をしている。 </◆ s2>

このような文対応は，人間の観察者 (たとえば，日英記事のスタイルを比較研究しているような人)にとっては価値があるが，文対応の結果を自然言語処理，たとえば，機械翻訳に利用しようとしている場合には，今のところは，有用性は限定されている．そのため，なるべく直訳同士にあるような文対応を抽出したいのであるが，このような状況から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スコアが必要である．

3 記事対応付けの方法

記事対応付けは，言語横断検索の枠組で行なう．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．そうした主な理由は，手元にある言語資源が日英方向の変換に便利だったからである．

3.1 日本語記事の英単語集合への変換

我々は，辞書引きに基づいて日本語記事を英単語集合に変換することにした．利用した日英辞書は，EDR日英対訳辞書，EDICT (一般的な日英対訳辞書)，ENAMDICT (固有名詞の日英対訳辞書)である[3]．これらの辞書の見出し語に対して，IPADIC (version 2.4.4) の品詞体系を付与し，茶筌[4] (version 2.2.8)の追加辞書として利用した．追加したエントリ数は，各辞書での重複を考慮しなければ，EDR日英対訳辞書が約 18万，EDICTが約 6万，ENAMDICTが約 22万である．

[3] http://www.csse.monash.edu.au/~jwb/edict.html
[4] http://chasen.aist-nara.ac.jp/index.html.ja

こうすることにより，茶筌の解析結果から容易に日英対訳辞書のエントリがアクセスできるようになる．たとえば，「あおぎ 見た月」は，追加辞書なしの状態では

あおぎ	あおぐ	動詞-自立
見	見る	動詞-自立
た	た	助動詞
月	月	名詞-一般

と形態素解析される (形態素情報の一部を省略)が，追加辞書ありの状態では

あおぎ見	あおぎ見る	動詞-自立
た	た	助動詞
月	月	名詞-一般

のように解析され，特に工夫をせずとも，複合語である「あおぎ 見る」の訳語として「look up」「face upwards」「look up to」「respcet」「admire」などが得られる．また，この方法によると，「くすの木台に行く」を形態素解析した場合のように，辞書にない単語に起因する解析誤りである「くす/の/木/台/に/行く」のようなものも「くすの木台/に/行く」として解析でき，かつ，「くすの木台」の訳語として「Kusunokidai」も容易に得られる．このように，IPADICを増強することにより，解析誤りを避けながら，容易に日英辞書の辞書引きができると共に，複合語の翻訳という言語横断検索において重要な作業も同時にできるため，この方法は有用である．

このようにして日本語の各単語 (もしくは複合語)において，その品詞が内容語 (主に名詞)に相当するものから英訳語を得て，そこから簡単なヒューリスティクスにより主辞を抽出し訳語としたが，このとき，全ての (相異なる)訳語の主辞を訳語として採用するとすると，英語の記事が質問として与えられたとき，質問中の思わぬ単語と一致することがありうるため，なるべく，訳語として正しいものだけを利用したい．そうするためには，訳語の曖昧性を解消すれば良いのだが，それを正確にするのは困難である．そのため，ここでは，ヒューリスティクスとして，まず，訳語の主辞の中から，より多くの訳語に含まれているようなものを優先し，次に，同順位のものについては，その訳語の主辞に対応する日本語単語を含む日本語記事の年と同年の英語記事において，その訳語の主辞を含む英語記事数 (document frequency, df)が多いような訳語の主辞を優先する[5]ことにした．そして，このヒューリスティクスにより優先付けられた上位2個のみを利用した．なお，dfが 0であるような訳語の主辞は，最初から，候補に含めない．

[5] たとえば，「今日」には「today」「nowadays」「this day」「present day」などが訳としてあるが，このうち，主辞だけをみると，「day」が一番多いので，まず，これを取る．次に，「nowadays」「today」の中から，dfが高いものを取る．

3.2 英語記事からの日本語記事の検索

一旦，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．そのためのソフトウェアとして，確率に基づくタームの重み(Robertson and Walker 1994)を利用した情報検索手法を実装した ruby-ir (内山 井佐原 2001)を利用した．

このとき，質問である英語記事 Q と日本語記事の変換結果 D の類似度は $\mathrm{BM25}(D,Q)$ である．ここで，

$$\mathrm{BM25}(D,Q)=\sum_{T\in Q} w^{(1)}\frac{(k_1+1)tf}{K+tf}\frac{(k_3+1)qtf}{k_3+qtf}$$

ただし，

T は Q に含まれる単語 (ターム)である．

$w^{(1)}$ は T の重みであり，$w^{(1)}=\log\frac{(N-n+0.5)}{(n+0.5)}$.

N は，検索対象の文書集合における全文書数である．ただし，検索対象は，質問である英語記事の日付の前後2日の範囲の日本語記事 (の英単語への変換結果)である．

n は，T を含む文書の数である．

$K=k_1((1-b)+b\frac{dl}{avdl})$ である．ただし，k_1，b，k_3 は経験的に定める定数であり，本研究では，$k_1=1$, $b=1, k_3=1000$ である．また，dl は，D の長さであり，$avdl$ は，文書集合における文書の長さの平均値である．ただし，文書の長さとは，その文書に含まれる単語の延べ数のことである．

tf は，D に含まれる T の数である．

qtf は，Q に含まれる T の数である．

4 文対応付けの方法

日英記事における文間の対応はDPマッチングで求めた (Gale and Church 1993; Utsuro, Ikeda, Yamane, Matsumoto, and Nagao 1994)．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には (Utsuro et al. 1994)を参照せよ．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，$\mathrm{SIM}(J,E)$ についてのみ述べる．

$$\mathrm{SIM}(J,E)=\frac{\mathrm{co}(J\cap E)+1}{|J|+|E|-2\mathrm{co}(J\cap E)+2}$$

である．ただし，$\mathrm{f}(x)$ を文 X における x の頻度とすると $|X|=\sum_{x\in X}\mathrm{f}(x)$ である．また，$\mathrm{co}(J\cap E)$ は，J 中の単語と E 中の単語との1対1対応を2言語辞書に基づきもとめた場合の集合を $J\cap E=\{(j,e)|j\in J, e\in E\}$ とすると，$\mathrm{co}(J\cap E)=\sum_{(j,e)}\min(\mathrm{f}(j),\mathrm{f}(e))$ である．なお，類似度については，詳細な比較検討はしていないが，実験結果は，この類似度の妥当性を支持するものである．

$J, E, J\cap E$ などは，以下のようにして求めた．まず，辞書引きにあたって，日本語文については，茶筌により形態素解析をした結果から，内容語および複合語を抽出し，英語文については，Brill's Tagger (Brill 1992)により品詞付けをし，基本形を WordNet [6] のライブラリを利用して求め，その結果から，内容語と複合語を抽出した．これらが J および E である．次に，$J\cap E$ については，ある (j,e) の組 $(j\in J\wedge e\in E)$ について，もし，j の訳語に e があるか，e の訳語に j がある場合には，(j,e) には対応の可能性があるとし，そのような全ての対応の可能性のなかから，訳語の曖昧性の低いほうから1対1に対応付けていった．すなわち，(j,e) の曖昧性として，j の訳語の数を採用し，それの小さいものから対応付けをしていくのだが，既に，(j,e) のどちらかでもが選ばれている対応はスキップする，という方法を採用した．なお，文対応付けに用いた2言語辞書は，EDR日英対訳辞書と EDR英日対訳辞書を統合して生成した日英および英日対訳辞書である．これらの辞書において，日英方向のエントリ数は約32万，英日方向のエントリ数は約37万である．

文対応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，$1\le n\le 6$ 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 (日本電子工業振興協会 2000)に適用することにより求めた．白書データには，18対の日英ファイルがあるが，そのうち，訳抜け (0対 n もしくは n 対0)の数が3以下の12ファイルを対象とした．これらのファイル対について，日本語文の平均数は413，英語文の平均数は495である．

このとき，再現率の平均は 0.982，適合率の平均は0.986である．ただし，

$$\text{再現率}=\frac{\text{プログラムの得た文対の中で正しい対の数}}{\text{正しい対の総数}}$$

$$\text{適合率}=\frac{\text{プログラムの得た文対の中で正しい対の数}}{\text{プログラムが推定した対の総数}}$$

ただし，1対 n の文対応からは，n 個の対が得られる．たとえば，文 J_1 と文 E_1, E_2, E_3 が対応しているとすると，得られる対は $(J_1,E_1),(J_1,E_2),(J_1,E_3)$ の3個である．

これより，このプログラムの精度は十分に高いと言える．そして，この精度の高さは，白書データにだけに限るのではなく，物語や小説などにも，ほぼそのまま，持ち越せることが，厳密な精度評価はしていないが，我々

[6] http://www.cogsci.princeton.edu/~wn/

のいくつかの経験から分かっている[7].

我々は，辞書のみに基づいて文対応付けをした．それに対して，(Utsuro et al. 1994)は，辞書情報に統計情報を組合せることにより，文対応の精度が向上すると述べている．しかし，我々のプログラムの精度は既に十分に高い[8]ので，統計情報は利用しなかった．

5 記事対応スコアと文対応スコア

3節において，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スコアとして，AVSIM(J,E) を定義する．これは，J と E との文対応を $\{(J_1,E_1),...,(J_m,E_m)\}$ としたとき，以下の式である．

$$\mathrm{AVSIM}(J,E)=\frac{\sum_{k=1}^{m}\mathrm{SIM}(J_k,E_k)}{m}$$

次に文対応スコアについて述べる．4節で述べたように，我々の文対応付けプログラムの精度は，原文と訳文とを対応付ける限りにおいては，高精度である．しかし，2節で述べたように，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，4節の方法で文対応付けをした場合には，適切な対応と共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，文対応のスコアとして文類似度だけでなく，記事対応スコア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 (集合) J_k と E_k との文対応スコアとして，

$$\mathrm{SntScore}(J_k,E_k)=\mathrm{AVSIM}(J,E)\times\mathrm{SIM}(J_k,E_k)$$

を定義した．このスコアは，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，記事対応スコアも高いような文対応を優先する．

6 記事対応付けの精度

6.1 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[9].

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，記事対応の良さの評価値であり，その基準は，Aは「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[10].「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節で述べたように，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．

我々の観察によれば，評価値が A/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，対応の良さにより類似度1位による記事対応をソートし，その上位のみを抽出すれば良い．

6.2 ソートした場合の記事対応の精度

記事対応のスコアとして，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，それぞれのスコアの降順により記事対応をソートし，評価値が A/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よりも適切なスコアであると判断した．

[7] http://www.crl.go.jp/jt/a132/members/mutiyama/snt-align/index.html より文対応付けされた小説等が得られる．

[8] 文対応付けの対象が異なるので厳密な比較はできないが，我々のプログラムの精度は (Utsuro et al. 1994)の最高精度よりも高い．この主な理由は，彼らの2言語辞書のエントリ数が約5万なのに対して，我々の辞書のエントリ数が30万を越えているからだと考えられる．

[9] 評価を記述する際には 1996-2001をメインとする．その理由は，今後とも The Daily Yomiuri には「本紙翻訳=Y/N」の情報が付くと考えられるので，1996-2001の精度評価の方が相対的に重要と考えられるからである．

[10] A,B,C,Dの判定は第1著者がした．判定については，1996-2001については，ダブルチェックをした．1996-2001についての初回の判定における各評価値の割合は，A=0.62, B=0.09, C=0.09, D=0.20である．したがって，同一評価者内においては判定結果は安定していると言える．なお，1996-2001については，更に，類似度1位の評価値がCかDの場合には，10位以内までを見て，A,Bがないかを探した場合の各評価値の割合は，A=0.62, B=0.15, C=0.05, D=0.18であるので，類似度1位のものとそれほど違わない．そのため，1989-1996については，類似度1位のもののみしか判定しなかった．

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

6.3 評価値と AVSIM

評価値 A,B,C,D と AVSIM との対応の良さを調べることを目的とし，表 1と同じデータに対して，各評価値となった記事対応について，AVSIMの統計量を求めた．それらを，1996-2001については表 3に，1989-1996については表 4に示す．これらの表において,「数」とは，その評価値であったような記事対応の数である．また,「平均」とは，そのような記事対応の AVSIMの平均値であり,「下限」および「上限」は，平均値の 95% 信頼区間の下限と上限である.「閾値」は，その評価値であるような記事対応と，次の評価値であるような記事対応とを分けるときに，どの AVSIMで区切れば良いかを示す．たとえば，表 3 では，A 判定と B判定とは AVSIMの値が 0.168により分かれる．この閾値は，線形判別分析により求めた値である．また,「有意差」の欄にある「**」と「*」は，それぞれ，その評価値と次の評価値とで平均値に差があるかを，Welch検定により片側検定したときに，その差が，1%と 5%水準で有意であることを示す．二つの表において，1989-1996 の Bと Cとの区分を除いては，全ての評価値において，各評価値と次の評価値とでは，平均値に有意な差があることがわかる．このことから，AVSIMは，各評価値を十分に明確に区切ることができると言える．なお，1989-1996では，Bと Cが分かれていないことについて，その理由を調べた．そうすると，実際，1989-1996 では，C だといっても，記述の重複が，1996-2001 の C と比べて，多いものが多かった．定性的には，1996-2001 の C は,「D ではない」という意味で C であり，1989-1996 の C は,「B かもしれない」という意味で C であった．

次に，1996-2001と 1989-1996とで，同じ評価値を与えられた記事対応の AVSIMの平均値に統計的に有意な差があるかを調べた．つまり，たとえば，表 3では，評価値 A の平均値は 0.193であり，表 4では，0.175である

表 3: AVSIMの統計量 (1996-2001)

評価値	数	下限	平均	上限	閾値	有意差
A	59	0.176	0.193	0.209	0.168	**
B	12	0.122	0.151	0.179	0.111	**
C	8	0.077	0.094	0.110	0.085	*
D	21	0.065	0.075	0.086		

表 4: AVSIMの統計量 (1989-1996)

評価値	数	下限	平均	上限	閾値	有意差
A	29	0.153	0.175	0.197	0.157	*
B	15	0.113	0.141	0.169	0.131	
C	8	0.092	0.123	0.154	0.097	**
D	48	0.076	0.082	0.088		

が，この二つの平均値の差が統計的に有意かどうかを両側検定による Welch検定により調べたところ，有意水準 5%においては，A,B,C,D いずれの評価値においても有意差はみられなかった．そのため，1996-2001と 1989-1996とで，同じ評価値の記事対応は，同じ程度の AVSIMであると判断した[11]．

最後に，表 3と表 4にある閾値[12]に基づいて，A,B,C,D であるような記事数を推定した結果を表 5に示す．表より，評価値が A/Bと推定される記事対応は，全体では，46738 (= 31495 + 15243)だけある．

表 5: 評価値と記事数の推定

	1996-2001	1989-1996	計
A	15491	16004	31495
B	9244	5999	15243
C	4944	10258	15202
D	5639	26825	32464
計	35318	59086	94404

7 記事対応付けの精度向上の可能性

6節で述べたように，AVSIMは，記事対応の良さを示す信頼性の高い尺度である．そのため，BM25の代り (もしくは重みつき和などで)，最初から AVSIMを利用して記事対応を求めれば，6.1節で述べた全体的な精度も向上すると考えられる．しかし，我々は，現時点では，BM25による類似度 1位の記事対応についてのみしか，AVSIMを求めていない．その理由は，10位以内などの比較的少しの記事をみただけでは記事対応精度に顕著な向上がないからであり，かつ，現時点での文対応プログ

[11]統計的に有意でなくとも平均値に差がある場合はありうる．

[12]表 3での閾値の丸めていない値は，0.1681076526, 0.111106681, 0.08531399165であり，表 4では， 0.1566618237, 0.130510963, 0.09692189387である．これらの値を実際には利用した

ラムの実行速度が遅いからである．今の文対応プログラムでは，一記事あたりの対応を取るために，数秒は掛る．そのため，一位同士の対応についてAVSIMを得るだけでも，9万4千記事程度なので，数日間は掛かる．したがって，たとえば，100位以内をみるだけでも，数100日間掛かることになる．これは非現実的である．しかし，今後，もっと高速の文対応プログラムを作り，それを利用することにより，より高精度な記事対応が得られるものと考えている．

また，今は，各英語記事について，その記事の日付の前後2日の範囲しか調べていないが，記事によっては，5日前のものが翻訳されているものがあった．このようなものまでカバーするためには，もっと広い範囲から対応候補記事を集める必要がある．

この2点は，システム全体を効率化しスケールアップすることにより達成可能なので，将来的には実現したい．

8 文対応付けの精度

2節で述べたように，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位の記事対応全てから得られる文対応全てを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[13]．このような文対応の数は，1989-1996と1996-2001を合せた全体で，約130万だけある．なお，ここでの文とは，日本語文については，簡単なプログラムにより，句点などで日本語記事を分割した結果であり，英語文については，MXTERMINATOR (Reynar and Ratnaparkhi 1997)に対して前処理と後処理を適用して英語記事を分割した結果である．

文対応のなかでは，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対応のなかで，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，約66万ある．

[13]文対応精度評価は，1989-1996と1996-2001とを分けずに行なう．その理由は以下の2点である．(1) まず，作成するコーパスでは，1989-2001全体から選んだ文対応のなかから良く対応していそうなもののみを抽出したい．そのためには，全体を評価した方が良い．(2) 記事対応の精度評価の結果から，同程度のAVSIMは，1989-1996と1996-2001とで同じ評価値に対応するので，SntScoreも1989-1996と1996-2001とで分ける必要はないと考えられる．

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[14]．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6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6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0.98である．

表 6: 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

表 7: 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

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6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．それを表7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．

1:nのデータは，単語などの対訳データを得るのに役立つだけでなく,「クリントン演説では直接日本に言及していないが、米側は日米同盟の重要性を繰り返し強調している。」「Clinton's address contained no direct references to Japan. However, the United States has long stressed the importance of the Japan-U.S. partnership. 」のように，長文分割の際の参考データとしても役立つものもある．また，文分割の誤りを回復させているものもある．

[14]評価は第1著者がした．ダブルチェックによると，初回の判定と2回目の判定とで100個あたり多くて2,3個程度のo/xの違いがあった．したがって，同一評価者内においては判定結果は安定していると言える．なお，1:nについてはダブルチェックはしていない．

9 関連研究

自動的に記事対応を得ることを目的とする研究はいくつかある．このうち，(Collier, Hirakawa, and Kumano 1998)は，言語横断検索に機械翻訳を利用した場合と辞書引きを利用した場合とを比較しており，再現率が高いとき(多くの記事対応を得たいとき)には，辞書引きの方が有利だとしている．我々も，表1のデータの1996-2001についてのみ，シャープ株式会社の機械翻訳支援システムを利用して精度評価をしてみたが，その結果は，統計的に有意ではないが，辞書引きの結果の精度の方が高かった[15]．これらのことから，辞書引きの方が記事対応を得るには適しているのではないかと考えられる．また，(Matsumoto and Tanaka 2002)は，日経産業新聞について，英語記事と日本語記事との対応付けをしていて，その精度は，97%と非常に高精度である．しかし，彼らは，同じ方法を，NHKの報道記事の対応付けに対しても利用しているが，その場合の精度は69.8%であり，彼らの方法が，全ての場合で高精度であるわけではないということも示している．そのため，彼らの方法を読売新聞の記事対応付けに利用した場合にも同様に高い精度が得られるかは明かではない．

従来の研究と我々の研究の主要な違いは，1節で述べたように3点ある．(1)まず，記事対応の評価スコアについて，我々は，DPによる文対応付けの結果を利用した信頼性の高いスコアを提案した．それに対して従来の研究は bag-of-words に基づいたスコアを利用している．なお，情報検索において，質問文と文書との類似度を求める際にDPを利用する方法が(Yamamoto, Yamamoto, Umemura, and Church 2000)により提案されているが，彼らの研究対象と我々の研究対象とは異なるし，かつ，DPの方法や，スコアの定義も異なる．(2)次に，我々は，記事対応の結果から文対応までを，実際に，大規模に得た．(高橋, 松尾, 古瀬 1999)は，記事対応の結果から文対応を得ることを構想してはいるが，実際に文対応を得ているわけではない．加えて，我々は，対応付けの結果が一般に研究利用できるようにしているが，これは日英対応コーパスとしては初めての試みである．(3)最後に，我々は，文対応付けプログラムを一般に利用可能にしたが，これは日英2言語コーパスに対応している文対応付けプログラムとしては，初めてのものである．

[15]ただし，このときには，英語記事を日本語に翻訳し，その翻訳結果を質問として日本語記事からなるデータベースを検索した．これは，本稿でこれまで説明してきた方法である，日本語記事を英語に変換する方法の逆であるので，厳密な比較ではない．

10 おわりに

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とから記事対応と文対応を得た．それらのなかで，比較的良質と推定されるものが，記事対応は約4万7千あり，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，現時点で一般に利用できる日英2言語コーパスとしては最大のものである．

更に，我々は，記事対応の評価スコアとして，文対応に基づいた信頼性の高いスコアAVSIMを提案した．現在の我々のシステムにおけるAVSIMの利用は，システムの速度の問題から，限られたものとなっているが，今後，システムを高速化することにより，もっと広い範囲でAVSIMを利用することができれば，精度の高い記事対応が更に多く得られるものと考えている．

参考文献

Brill, E. (1992). "A Simple Rule-Based Part of Speech Tagger." In *ANLP-92*, pp. 152–155.

Collier, N., Hirakawa, H., and Kumano, A. (1998). "Machine Translation vs. Dictionary Term Translation – a Comparison for English-Japanese News Article Alignment." In *COLING-ACL'98*, pp. 263–267.

Gale, W. A. and Church, K. W. (1993). "A Program for Aligning Sentences in Bilingual Corpora." *Computational Linguistics*, **19** (1), 75–102.

Matsumoto, K. and Tanaka, H. (2002). "Automatic Alignment of Japanese and English Newspaper Articles using an MT System and a Bilingual Company Name Dictionary." In *LREC-2002*, pp. 480–484.

Reynar, J. C. and Ratnaparkhi, A. (1997). "A Maximum Entropy Approach to Identifying Sentence Boundaries." In *ANLP-97*.

Robertson, S. E. and Walker, S. (1994). "Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval." In *SIGIR'94*, pp. 232–241.

Utsuro, T., Ikeda, H., Yamane, M., Matsumoto, Y., and Nagao, M. (1994). "Bilingual Text Matching using Bilingual Dictionary and Statistics." In *COLING'94*.

Yamamoto, E., Yamamoto, M., Umemura, K., and Church, K. W. (2000). "Dynamic Programming: A Method for Taking Advantage of Technical Terminology in Japanese Documents." In *IRAL-2000*, pp. 125–132.

内山将夫 井佐原均 (2001). "情報検索パッケージの実装." 情報処理学会研究報告, FI-63-8, pp.57–64.

高橋大和, 松尾義博, 古瀬蔵 (1999). "新聞記事における日英対応コーパスの自動構築." 言語処理学会第5回年次大会発表論文集, pp. 181–184.

日本電子工業振興協会 (2000). 自然言語処理システムに関する調査報告書.