日英新聞記事コーパスにおける対応付けの精度評価

松村 繁男† 絹川 博之† 山田 剛一† 中川 裕志‡
Shigeo Matsumura Hiroshi Kinukawa Koichi Yamada Hiroshi Nakagawa

１．はじめに

現在、日英新聞記事コーパスにおいて自動的に記事対応を得ることを目的とした研究はいくつかある。通信総合研究所の内山らは、読売新聞と The Daily Yomiuri を基にし、記事対応を１対１とし、自動的に記事対応を得ている。しかし、１件の英文記事が１件の邦文記事に必ずしも対応するのではなく、複数の邦文記事を合成したものであることが多い。また、対応する英文記事の掲載は、ほとんどが基となる邦文記事の掲載日の翌日であるが、時々遅くなることもある。

英文記事と邦文記事の記事対応は１対多であるということを検証するために、我々も読売新聞社の「読売新聞記事データ」を使用し、一つの英文記事に対し、複数の邦文記事を対応付ける実験を行った。

本稿では、この対応付けの結果を基に、通信総合研究所の内山らの結果[1]との比較評価実験を行った。比較評価には平均精度を用いた。

以下では、まず、日英新聞記事コーパスの対応付けの方法を述べ、次に精度評価方法と比較結果を示し、最後に結論を述べる。

２．日英新聞記事の対応付け方法

記事の対応付けは、内山らと同様に英文記事を質問とし、それに類似する記事を邦文記事データベースから検索するという言語横断検索の枠組みで行う。これにより与えられた英文記事に対応する邦文記事を見つけることができるようになる。

このとき内山らは、邦文記事を英単語集合に変換して日英新聞記事を対応付けているのに対し、我々は英文記事を翻訳記事集合に変換して両記事を対応付けた。この対応付けに際して、我々は汎用連想計算エンジン(GETA) [2]を用いた。

２.１ 邦文記事集合からWAMの生成

WAM とは、GETA で扱う行列のインスタンスである。行は 1 件 1 件の記事を示し、列は各記事に出現する単語を示し、行列の要素は単語の出現回数を示すものである。WAM を作成するには、記事ごとに茶筌[3]で形態素解析を行い、その結果から品詞が名詞と未知語である単語の出現回数をカウントの対象とした。

２.２ 英文記事集合からWAMの生成

英文記事集合から WAM を生成するには、前述したとおり、英文記事集合をまず翻訳記事集合に変換する必要がある。我々は、機械翻訳ソフト「IBM インターネット翻訳の王様バイリンガル Version5」のテキスト翻訳機能を用いて英文記事を翻訳記事集合に変換した。英日翻訳エンジンでの辞書の語数は約 40 万語である。

次に翻訳記事集合を基に WAM を生成していく。生成は邦文記事同様、記事ごとに「茶筌 Version2.2.9」で形態素解析を行い、単語をカウントしていく。カウントの対象となる単語は邦文記事集合と同様である。しかし翻訳記事集合の中で、翻訳されていない英単語は以下の方式で出現回数をカウントする。

(1) ローマ字かな変換ライブラリ[4]を用いて音訳を行う。(例:matsumura→マツムラ)
(2) REMTOSS 社の姓名辞書を用いて音訳されたカナ列をその読み方の漢字列に変換し、候補の漢字列をすべて出力する。(例：マツムラ→松村、末村 etc)
(3) 出力された漢字列および音訳されたカナ列を単語として出現回数をカウントする。

２.３ 英文記事から類似邦文記事の対応付け

2.1 節、2.2 節で生成した WAM を入力として、GETA を用いて英文記事から類似する邦文記事を検索することができる。GETA は、記事を特徴付ける特徴語を基に類似度を計算し、上位から出力してくれるソフトである。

今回はこの特徴語の数を 50 として計算し、表示される類似記事のうち上位 20 個を取り出した。

３．対応付け結果への補正

前節の日英新聞記事の対応付けの結果には、これは英文記事１つに対し、邦文記事集合全体から対応付けを行った結果なので雑音が多く含まれている。本節では、それぞれの対応付け結果に対する補正方法について述べる。

(1) まず、出力された対応付け結果に対して日付によるフィルタリングを行う。日英の対応する記事を検討した結果、邦文記事が掲載された後に対応する英文記事が掲載されるというパターンが多かったので、対応付け結果として出力された類似記事の中で、対応元となる英文記事の日付の 10 日前から 3 日後までの間に掲載された邦文記事を取得対象とし、この間に掲載されていない邦文記事は取得対象外とした。
(2) 次に、取得対象とした邦文記事に対して日付による類似度の補正を行う。GETA によって出力された類似度のみによる場合では、正解集合にばらつきがでてしまう。そこで、この類似度にさらに日付による重みを加えることにした。その計算式を(式１)で示す。これは出力された類似度 SIM に日付による重み定数 C をかけるものである。

$$MULSIM = SIM \times C \quad \text{(式 1)}$$

Precision evaluation of matching in Japanese-English News article corpus

†東京電機大学工学部
(School of Engineering, Tokyo Denki University)
‡東京大学
(The University of Tokyo)

重み定数は、基となる英文記事の日付と類似記事の日付との距離によって決定する。日英の対応する記事を検討した結果、類似記事は英文記事の日付の前日に掲載されていることが多いことがわかった。したがって英文記事の日付の前日、つまり差が-1 の C の値を一番大きくしている。英文記事と日付の一致する類似記事は、前日のものよりは重要ではないが、2 日前の記事よりは重要であると考え、一致する日付の重みは前日と 2 日前の間と設定した。重み定数 C を表 1 に示す。

表 1　スコアのフィルタリング詳細

差	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1
C	1	2	3	4	5	6	7	8	9	10
差	0	1	2	3						
C	9.5	7	6	5						

この結果を基に検索結果をソートし、上位 10 個を類似記事として出力する。ここで類似度 1 位の記事だけでなく、複数の類似記事を出力する理由は対応記事が常に 1 位として出力されるわけではないということと、1 節で述べた理由からである。

4．記事対応付けの精度

記事対応付けの精度の評価方法について述べる。評価は平均精度[5]を用いることとした。平均精度の数学的定義は(式 2)で示す。

$$v = \frac{1}{\sum_{i=1}^{N} x_i} \sum_{i=1}^{N} \left[\frac{x_i}{i} \left(1 + \sum_{k=1}^{i-1} x_k \right) \right] \quad \text{（式 2 ）}$$

なお、N は出力文書の総数、X_i 出力順第 i 位の文書の適合／不適合の状態を示す変数とする。適合／不適合の判定は 2 値とし、適合ならば X_i = 1、不適合ならば X_i = 0 とおく。

4.1　平均精度算出データ

平均精度算出データは以下のものとする。

(1) まず当該英文記事が、ある邦文記事の翻訳により作成されたことを示す「本誌翻訳=Y」となっている記事を評価対象として 2 つのデータを用意する。
  (a) 1 つ目は、対象となる記事を 1 対 1 対応として捉えたものである。これは、内山らの検索実験で用いたものである。このデータ名を「1 対 1 対応平均精度」とする。
  (b) 2 つ目は対象となる記事を 1 対多対応として捉えたものである。このデータ名を「1 対多対応平均精度」とする。

(2) 3 つ目は全英文記事を評価対象とし、1 対多対応として捉えたものである。このデータ名を「全データ平均精度」とする。

また、類似記事の正解データは内山らの示した正解データを基準とし、1 対多対応の場合は、内山らの正解データに我々が正解であると考えたものを加えた。

4.2　日付による重み付けの精度評価

表 2 にそれぞれの精度評価結果を示す。

精度評価値は、GETA の対応付けに対して日付によるフィルタリングと類似度の補正を行う前と後の 2 通りを示す。なお、補正前のデータは無作為に抽出した 25 件の記事であり、補正後のデータは、補正前の 25 件の記事に無作為に抽出した 175 件の記事を加えたものである。

表 2　記事対応の精度

	評価値	
評価項目	重み付け前	重み付け後
評価データ量	25 件	25+175 件
1 対 1 対応平均精度	81.67%	86.75%
1 対多対応平均精度	76.84%	93.30%
全データ平均精度	78.75%	92.50%

5．おわりに

(1) まず精度評価の結果から次の事が得られた。
  (a) 日付による補正を加えることにより、内山らのものより精度の高い結果を得られる可能性を示すことができた。
  (b) 日英の新聞記事を対応付ける際には、語の頻度で重みをつけるだけではなく、日付による重み補正をすることが重要であるといえそうである。
  (c) 記事対応についても半数以上が 1 対多対応であった。したがって記事の対応付けは 1 対多と考えて行うほうがよいようである。

(2) そして、今後の課題として次のことが挙げられる
  (a) 我々の精度評価データ量は、現時点では少ない。今後、大量のデータを用いて評価し、結果の信頼性を高くする必要がある。
  (b) これらの対応付け結果を基に日英両文章の対応付け方式を考えていく。そして、記事対応が 1 対多であるということを用い、得られた文章の対応を基に複数文書要約の方式を研究していく予定である。

参考文献

[1] 内山将夫　井佐原均(2002):“日英新聞記事の対応付けと精度評価.”情報処理学会研究報告「情報学基礎」, No.068, pp.15-22.

[2]（株）日立製作所, 東京工業大学, 北陸先端科学技術大学院大学, 文部省国文学研究資料館(2002):“汎用連想計算エンジンの開発と大規模文書分析への応用.”

[3] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸(2000):“形態素解析システム茶筌.”http://chasen.aist-nara.ac.jp/

[4] 増井俊之(2001)：“ローマ字かな変換ライブラリ.”http://www.csl.sony.co.jp/person/masui/OpenPOBox/romakana/

[5] 岸田和明(2001):“検索実験における評価指標としての平均精度の性質.”情報処理学会トランザクション「データベース」, Vol.43 No.SIG02, pp.97-104.