ＤＢから抽出した日英新聞記事の自動対応付け*

高橋大和　白井諭　藤波進　池原悟

NTT　コミュニケーション科学研究所

上田洋美　松島英之

NTT　アドバンスドテクノロジ(株)

1. はじめに

機械翻訳などの自然言語処理技術を研究する上で、大量の対訳コーパスは非常に有用であるが、収集するのは非常に難しい。従来、日英対訳コーパスとして、1. 英文マニュアルを日本語へ翻訳したもの 2. 英和・和英辞書の用例 3. 独自に開発した小規模のもの などが挙げられるが、大量の一般的なデータの収集は困難である、という問題点がある。

これに対して、新聞記事を対象とすれば継続的にデータを収集することができる利点があるが、従来、対応付けが容易でないと考えられてきた。しかし、日英のように言語のギャップが大きい場合でも、記事内容に着目すれば、基本的に一致すると考えられる。

例えば、日本経済新聞社が提供しているテレコンＤＢから取得した日英記事を比較検討した例では、部分対応を含めると、ほとんどの英文に内容的には対応する日本文があり、そのうち5割は格要素などの対応もとることができることが報告されている[1]。

対訳コーパスを構築するためには、1. 日英記事対応, 2. 日英文対応 を行う必要がある。1. に関して、数値と名詞をキーワードとして統計的に対応を行う方法が提案されている[2]が、この手法では、対応の採否の判定基準の決定が困難であるという問題があり、最終的には人手による確認が必要であると考えられる。

そこで、本稿では自動的な記事対応付けの手法の確立をめざし、対応項目数により採否の判定を行う評価アルゴリズムを提案する。

２．記事の特徴

本稿で対象とした記事は、日本経済新聞社が有料情報サービスとして提供しているテレコンＤＢから、電話回線経由のパソコン通信により取り寄せることができる。日本文記事は、日経テレコンＢＩＺに収録されている日経四紙（日本経済新聞, 日経産業新聞, 日経流通新聞, 日経金融新聞）を対象とし、英文記事は、Nikkei Telecom Japan News & Retrieval より、日経四紙の速報訳として提供されている記事を対象とした。

１９９４年１１月１日から１０日までの記事数の例を表1に示す。日本文記事は英文記事に比べ7倍近い記事数が収録されていることがわかる。したがって、英文記事を基準として対応する日本文記事を得るほうが効率がよい。

対訳テキストにおける文の対応付けには、統計的情報や対訳辞書の利用が有効であることが知られている[3][4][5]が、数百記事の中から目的とする記事を見つけるために応用することは困難であると考えられる。一般的には、記事を特定する情報として、ＤＢ検索ではキーワードを利用する。この方法を応用した場合、対応する記事の検索に英日の対訳キーワード辞書を利用して、日本文記事の検索を行う。しかし、この場合は対応付けの品質は対訳辞書の品質に大きく左右される。

そこで、通常はキーワードとして利用されていない数値情報に着目する。数値情報は、日英記事のどちらにおいても、記事からの抽出が容易であり、対応も取り易い。よって、記事を特定する情報として有用であると考えられる。例として、数値情報を含む記事数を表1に示す。数値を含む記

* Automatically Aligning Newspaper Articles from Databases

事は日本文で約 80%、英文で約 90%あることがわかる。

数値情報と名詞をキーワードとする方法としては、数値情報だけで約 50%、名詞キーワードを含めると 80%の対応付けを行えることが知られている[2]。この方法は、キーワードの出現回数から計算した稀少度を用いて対応の評価を行う。しかし、対応記事の採否の判定基準となる稀少度は記事の数やキーワードの出現回数に左右されるため、その基準を決めるのは困難である。そのため、最終的にはアナリストによる評価を行わなければ対訳データとして使うには精度が低いと考えられる。

そこで、キーワードの出現回数に代えて、対応項目数に着目する。数値の組は、記事の個別性を表す特徴的な情報として考えられるためである。本稿では、対応項目数により対応記事の採否を判定する日英記事対応アルゴリズムを提案する。

表1　日本文記事と英文記事の量(1994 年 11 月分)

日付	曜日	日本文記事			英文記事		
		記事数	平均字数	含数値	記事数	平均Bytes	含数値
1	火	1001	410.2	843	137	792.9	125
2	水	842	408.5	703	120	775.4	111
3	木	485	396.4	360	39	830.6	34
4	金	738	449.4	647	97	803.0	85
5	土	504	367.7	358	29	905.4	26
6	日	167	588.4	127	12	829.2	8
7	月	629	541.5	519	128	758.8	116
8	火	929	451.9	671	152	754.5	132
9	水	819	410.8	641	146	782.3	131
10	木	941	471.7	756	151	777.3	142
合計		7055	437.7	5625	1011	783.0	910

3．記事対応付けアルゴリズム

対応する記事を発見する際の問題として、検定すべき候補記事はかなり多数にのぼることが表1からわかる。そこで、本稿では字面処理程度の浅い解析による方法を中心に考察した。具体的には、英文記事の本文に含まれる数値と名詞をキーワードとして対応付けを行う。以下に、キーワードの抽出アルゴリズムを述べる。

3．1　英文記事

記事は 1 日分を対象として、その範囲内に含まれる数値およびキーワードをすべて抽出する。

3．1．1　数値キーワードの切り出し

1. 本文の数値を数値リストとして切り出す。小数点や次の単位が続いている場合はそれらを含めて切り出す。ここで扱う単位は、出現頻度が高く、日本文においても切り出しやすい単位のみを扱う。
   例：dollar,yen, %, trillion, billion, million
2. “trillion, billion, million” を含む数値を数字列に正規化する。
   例：4.3 trillion dollar　→
   4300000000000 dollar
3. 数値リストは記事単位で項目の重複がないように、重複した数値項目を削除する。
4. 数値リストをソートし、項目の出現回数を数え、数値項目重みデータとする。
5. 数値リストに、数値項目重みデータの重みを付加する。これを、**英文記事重み付き数値リスト**とする。

3．1．2　名詞キーワードの切り出し

1. 本文と見出しから名詞と推定される単語を名詞リストとして切り出す。以下の条件を満たす単語は一単語とみなす。
   ・大文字を含む単語列
   ・空白,コロン,ピリオド,カンマ,シングルクォートを単語間に含む単語列
2. 名詞リストの項目をキーとして、対訳辞書を検索する。日本語訳があった場合、名詞リストに日英の対として追加する。訳がなかった場合は項目を削除する。
3. 名詞リストは、記事単位で日本語訳の重複がないように、重複した単語項目を削除する。これを、**英文記事キーワードリスト**とする。

3．2　日本文記事

英文 1 日分に対して、日本文は英文記事の日付前後 1 日を含めた 3 日分を対象として対応付けを行う。これは、白井らによって報告されているように[2]、対応する英文記事と日本文記事の日付が同じ記事は 63.5%、英文記事が 1 日遅いものは 30.6%のほか、1 日早いものが 5.9%含まれているからである。

この 3 日分の日本文記事から数値をすべて抽出する。なお、日本文記事から固有名詞を直接抽

出することは容易でないため、対応付けの対象範囲として、見出し文およびリード文（第一段落）を抽出する。

3．2．1　数値キーワードの切り出し

1. 数値を切り出す条件は、数字、漢数字の連鎖、また、小数点、「ドル」、「円」、「銭」、「%」が続いている場合、これらを含めて一単語として扱う。
   例：　三十五円四十銭、五〇．八%
2. 数値を数字列に正規化する。
   例：　三十五円四十銭 → 35.40 yen
3. 数値リストは記事単位で項目の重複がないように、重複した数値項目を削除する。
4. 数値リストをソートし、項目の出現回数を数え、数値項目重みデータとする。
5. 数値リストに、数値項目重みデータの重みを付加する。これを、**日本文記事重み付き数値リスト**とする。

3．2．2　リード文の抽出

1. 日本文記事三日分のタイトルと第一段落をリード文として切り出す。これを、**日本文記事リード文リスト**とする。

3．3　対応付け

次のように、数値のみによる対応付け，英文から得たキーワードによる対応付けの２つの方法を試みた。

3．3．1　数値の記事対応

英文記事重み付き数値リストと日本文記事重み付き数値リストの対応付けを行う。対応付けは、正規化された数値のマッチングによって行われる。

↓

この結果を、数値対応付けリストとする。

3．3．2　キーワードの記事対応

英文記事キーワードリストと日本文記事リード文リストの対応付けを行う。対応付けは、英文記事名詞リストの名詞が日本文記事リード文リストのリード文に含まれているかどうかにより行われる。

↓

この結果を名詞キーワード対応付けリストとする。

4．対応付けの評価と考察

前節で述べたアルゴリズムで得られた対応付けリストの評価実験を行った。実験１として、稀少度が最も高くなる日本文記事:英文記事での数値の出現回数が１対１の対応を調べた。結果を表2に示す。ただし、項目の個数が同数の場合は、対応候補を決定しない。この場合は、稀少度を用いて、相対的に判定するためである。ただし、この実験では、稀少度による評価は行わない。

［実験１］
数値対応付けリストにおいて、日英の出現頻度が１対１の数値を含む記事対応で個数が多いもの。

表２　1対1対応（1994年11月2日～9日分）

日付	2日	3日	4日	5日	6日	7日	8日	9日
1個	15	4	14	5	0	18	9	23
	(20)	(5)	(17)	(6)	(0)	(22)	(14)	(25)
2個以上	9	7	22	5	2	21	24	17
	(9)	(7)	(22)	(5)	(2)	(21)	(24)	(17)

正解数
（記事数）

結果として、１対１の出現回数による評価では、対応項目の個数が１個の場合は、平均で約 70%程の正解率であり、２個以上の組み合わせがある場合は、100%の正解率が得られる。

これより、出現回数による評価より、対応項目個数に着目して評価を行った方がよいと考えられる。そこで、以下の実験２を行った。

［実験２］
数値対応付けリストにおいて、対応する数値項目の個数が多いもの。ただし、項目の個数が同数の場合は、対応候補を決定しない。

結果を表3に示す。

表３のデータを参考に、実験により候補の採否判定条件を検討した。その結果、候補が１つの場合は対応する個数が３個以上、複数候補がある場合は、第一候補と第二候補の差が２個以上の条件のとき、100%の正解率が得られた。

表3 対応項目数と正解数（1994年11月2日〜9日分）

個数	2個	3個	4個	5個	6個以上	合計
2日	7/0 (8/0)	0/6 (0/7)	2/9 (2/10)	1/5 (1/5)	0/18 (0/18)	10/38 (11/40)
3日	2/0 (2/0)	3/3 (3/3)	1/0 (1/0)	0/2 (0/3)	1/7 (1/7)	7/12 (7/13)
4日	2/0 (2/0)	4/3 (4/6)	2/10 (2/10)	1/1 (1/1)	2/21 (2/21)	11/35 (11/38)
5日	2/0 (2/0)	0/4 (0/4)	0/2 (0/2)	0/2 (0/2)	0/6 (0/6)	2/14 (2/14)
6日	0/0 (0/0)	0/1 (0/1)	0/1 (0/1)	0/0 (0/0)	0/3 (0/3)	0/5 (0/5)
7日	6/0 (7/0)	0/7 (0/9)	2/5 (2/6)	2/6 (2/6)	3/26 (3/26)	13/44 (14/47)
8日	7/0 (7/0)	5/7 (5/7)	2/13 (2/13)	0/6 (0/6)	0/23 (0/23)	14/49 (14/49)
9日	1/0 (3/0)	3/8 (3/10)	1/11 (1/11)	0/7 (0/7)	2/30 (2/30)	7/56 (9/58)

候補が1つの正解数／候補が複数の正解数

（候補が一つの対応記事数／候補が複数の対応記事数）

最後に、実験2の条件で、数値対応付けリストと名詞キーワード対応付けリストをマージしたデータに対して評価した結果と稀少度による評価の結果を比較した。これを表4に示す。

表4 数値とキーワードによる対応付けと稀少度による対応付けの比較（1994年11月2日〜9日分）

日付	2日	3日	4日	5日	6日	7日	8日	9日
*1	35	17	35	12	2	37	51	51
*2	13	4	12	2	1	12	9	5
合計	48	21	47	14	3	49	60	56

*1 稀少度による方法と同じ対応記事の数

*2 新たに見つかった対応記事の数

表4により、提案した方法により稀少度による対応付けでは得られなかった対応が得られることがわかる。

5. 問題点と今後の改良

現在、わかっている問題点として、名詞キーワードの拡張が必要である。解決法としては、数値による対応付けを行った後、名詞キーワードの対応を人手で行うことが挙げられる。

また、対応する個数が同数の場合、日本文記事の候補を決定するには稀少度による相対的評価を行うのがよいと考えられるが、稀少度の計算において、数値の単位による評価値の加減が必要と考えられる。特に、数値で出現頻度の多いものは、年度、記事の前後の日付がある。

本手法では、名詞キーワードを数値項目と区別せず、マッチングの個数だけで評価している。この点も、実験により決定していきたい。

6. おわりに

本稿では、並立するデータベースから収集した日英新聞記事を、浅い解析で得られる数値情報と名詞キーワードに着目して対応項目数を評価し、対応記事を自動的にかつ高い精度で得る方法を提案した。日英新聞記事8日分に対する対応付け実験の結果、対応項目が2個以上の対応記事候補が1つの時は対応項目が3個以上、対応記事候補が複数ある場合は第一候補と第二候補の差が2個以上の第一候補を対応記事候補とする、という条件で、1日平均約 41.5%の記事対応を正解率100%で得ることができた。今後は対応項目が同数の候補記事の評価方法と重みづけを実験により検討する。

本手法により、大量の日英の対訳記事を収集することが可能になり、新語や専門用語の対訳の収集、対訳表現の抽出など、辞書の整備や翻訳表現調査の効率化が図れると考えられる。また、白井[6]に提案されている文対応の方法を実験し、対応情報をSGMLタグとして構造化し[7]、継続的な大量の対訳コーパスの構築を目指す予定である。

参考文献

[1] 白井,藤波,池原,上田,井上：新聞記事日英対訳コーパスの構築(1)－基本構想と検討課題－，電気関係学会九州支部第48回連合大会(1995)

[2] 白井,上田,阿部,藤波,池原：新聞記事日英対訳コーパスの構築(2)－並立DBから取得した記事の対応付け－，電気関係学会九州支部第48回連合大会(1995)

[3] P.F.Brown,J.C.Lai & R.L.Merce : Aligning sentences in parallel corpra,29th ACL(1991)

[4] S.F.Chen : Aligning sentences in parallel corpra,31st ACL(1993)

[5] T.Utsro, H.Ikeda, M.Yamane, Y.Matsumoto & M. Nagano : Bilingual text matching using bilingual dictionary and statistics, 15th COLING(1994)

[6] 白井,松尾,瀬下,藤波,池原：新聞記事日英対訳コーパスの構築(3)－記事の特徴分析と文の対応関係の検討－，電気関係学会九州支部第48回連合大会(1995)

[7] F.Bond, Y.Takahashi, S.Yamada, M.Nisigaki : Still tagging an aligned Japanese/English corpus, 言語処理学会第2回年次大会(1996)