日英新聞記事の記事対応コーパス自動作成

高橋 大和　　白井 諭　　大山芳史
NTT　コミュニケーション科学研究所
渡邊いづみ　　上田 洋美
NTT　アドバンステクノロジ(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。しかし、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。

しかし、新聞記事を対象として、再現率よりも適合率を重視し、数値をキーワードとして利用することにより、記事対応を行うことができることが報告されている[1]。

これは、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ためである[2]。

本稿では自動的な記事対応付けの手法の確立をめざし、数値キーワードと併用して名詞キーワードを利用することによる効果と、より長い期間の記事データに対して適応した場合の結果とその問題点について報告する。

2.数値による記事対応

ここでは、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）を日本文記事として、また、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英文記事として実験を行った。

1994年11月2日から9日までの8日間の英文記事に数値による記事対応を行い、第一候補と第二候補の対応項目数の差が2個以上の時、正しい対応記事とみなす、という条件で、表1のような結果を得た。

この結果を基に、人手により日英の対訳名詞辞書を作成し、その効果を確認した。

2日	3日	4日	5日	6日	7日	8日	9日	合計
30	14	39	8	2	39	44	46	222

表1　数値による記事対応

3.数値と名詞キーワードによる記事対応付け

本稿では、多数の候補記事に対して効率よく対応項目を見つけるために、字面処理程度の浅い解析による方法を用いる。

3.1.英文記事からの名詞キーワードの抽出

記事は1日分を対象として、その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。抽出条件を以下に示す。また、こ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
  例1：NTT Communication Science Lab.
  例2：SL-enhanced Intel i486SX
- 大文字を含む単語列の所有格に大文字を含む単語がある場合
  例1：Japan Federation of Employers' Associations
  例2：International Standardization Organization's ISO9001
- “of”, “&” を大文字を含む単語間に挟んでいる場合
  例1：Social Democratic Party of Japan
  例2：Mitsubishi Trust & Banking Corp.
- 大文字を含む単語列の所有格に大文字を含む単語列が連接していない時、また、”of” の後ろに大文字を含む単語が連接しない時は、そこまでで切り出す。
  例1：NTT's line → NTT
  例2：Bank of city → Bank

・“The” は単語列に含まない。これは、文の途中では小文字になり、切り出し単語が増えてしまうためである。

例: The U.S. → U.S.

3.2 対訳リストの作成

抽出した名詞キーワード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、対訳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。

対訳リストは、記事単位で日本語訳の重複がないように、重複する単語があった場合は削除する。

3.3.日本文記事

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。これに対して、対訳リストのマッチングを行う。

3.4.記事対応付け

3.2.節により切り出した対訳リストを基に、数値による記事対応で対応が付いた記事から、対訳辞書を人手により作成した。内容を表2に示す。

内容	項目数
企業名	588
製品名	107
人名	23
地名・その他	136
合計	1344

表2 英日対訳辞書

新聞記事を対象としているため、企業名が一番多く出現する。表6から、実際のデータにおける出現率も、企業名が一番多いことがわかる。

この辞書を用いて、数値による記事対応と名詞による記事対応を併用して、記事対応実験を行った。

ここで、記事単位で見た時、名詞キーワードによる記事対応でマッチングした対訳語の一部分を含んでいる別の対訳語があった場合、長い単語のみ残し短い対訳語を削除する。

例：Tokyo　東京

Bank of Tokyo　東京銀行

この二つの対応対訳語があった時は、

“Bank of Tokyo　東京銀行” のみ残す。

また、数値を含む対訳語がマッチングした場合、数値対応と重複することになる。そこで、数値対応とマージする時に、数値対応の該当する項目を削除する。これは、名詞の方がより長いマッチングを行っているため、信頼性が高いと考えるからある。

3.5.結果

対応記事の第一候補と第二候補の対応項目数の差が2個以上という条件で、数値による対応付けと名詞キーワードによる対応付けをマージしたデータに対して評価を行った。結果を表3に示す。

日付	2日	3日	4日	5日	6日	7日	8日	9日	合計
*1	46	17	44	13	4	52	56	56	288
*2	30	14	39	8	2	39	44	46	222

表3 数値とキーワードを併用 (1994/11/2-9)

*1:数値とキーワードによる正解対応記事数

*2:数値のみの正解対応記事数

これより、数値と名詞キーワードを併用した場合、66記事（約29.7%増加）の新しい対応を得ることができることがわかり、効果を確認することができた。ここで、新しく選られた記事対応から、さらに新しい名詞対訳語を人手により抽出した。

結果を表4に示す。

内容	項目数
企業名	8
地名・その他	3
合計	11

表4 新規に得た英日対訳

新規の対訳を加え、同じ条件で対応づけを行った。結果は、11/02の対応記事が一つ増えたのみだった。対訳辞書の効果は高いが、再帰的な適応の効果は小さい。新聞記事の特徴として、記事の示すトピックはある会社のことに特定されるためであると考えられる。

4.一ヶ月分のデータに対する記事対応

ここで、提案した記事対応を評価するため、一ヶ月分のデータに対して、記事対応を行った。対象とする新聞記事データは1995年8月1日から31日までの英文、日本文記事とした。ここで、記事対応は、前後の日付に対しても行うため、日本文は7月31日から9月1日分までを用意した。

記事対応の結果を表5に示す。

日付	8/1	8/2	8/3	8/4
*1	48/48	41/41	58/58	39/39
*2	41/41	38/38	55/55	31/31
8/5	8/6	8/7	8/8	8/9
8/10	6/6	37/37	45/45	49/50
7/9	6/6	31/31	42/43	43/44
8/10	8/11	8/12	8/13	8/14
47/48	37/39	11/11	1/1	20/20
42/44	33/34	8/8	1/1	16/16
8/15	8/16	8/17	8/18	8/19
50/50	46/49	46/47	40/40	14/14
43/43	40/42	43/43	37/37	10/10
8/20	8/21	8/22	8/23	8/24
9/9	31/31	41/41	44/45	51/53
7/7	27/27	35/35	38/39	39/41
8/25	8/26	8/27	8/28	8/29
38/39	16/16	6/6	42/42	43/44
33/34	13/13	6/6	34/34	33/34
8/30	8/31	合計		
41/41	49/49	1054／1069		
33/33	40/40	905／919		

表5　記事対応結果(1995/8/1-31)

*1：数値と名詞キーワードによる記事対応(正解数／総数)
*2：数値のみによる記事対応(正解数／総数)

結果として、明らかに間違った記事が選択されたものが6記事、最近の動向の記事として紹介された記事の極一部が対応しているものが１１記事であった。

正解率は、数値のみで９８．５%、併用の場合、９８．６%であった。東証の記事における数字の偶然の一致、また、英文が長く日本文が短い、もしくはその逆という形で間違っている。

また、対訳辞書を用いることにより、１４９記事(１６．５%)の新しい対応付けを得ることができている。このことから、辞書による効果は高いと考えられる。

5.おわりに

今回の実験により、対訳辞書を用いた名詞キーワードによる対訳付けの効果を確認できた。８日間分のデータから抽出した対訳語句においても、効果があることが分かった。問題としては、効率よく名詞キーワードを収集する方法が必要である点である。これには、カタカナ語と漢字の読みによる自動的な単語対応づけ[3]を適用し、実験を行いたい。

また、候補記事の第一候補と第二候補の項目数の差が二個という条件では、まだ不十分な面もある事が分かった。キーワードの長さや数値の単位による評価値の加減が必要と考えられる。また、差が小さい場合にも、対応が正しいと考えられる条件を見つけていきたい。

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[4]に提案されている文対応の方法を実験をすすめ、文対応とのバランスを考えながら、記事対応の評価をしていきたい。また、記事対応を行ったデータはSGMLタグとして構造化し[5]、継続的な対訳コーパスの構築を目指す予定である。

参考文献

[1] 高橋,白井,藤波,池原,上田,松島：ＤＢから抽出した日英新聞記事の自動対応づけ,言語処理学会第２回年次大会(1996)

[2] 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)ー基本構想と検討課題ー，電気関係学会九州支部第48回連合大会(1995)

[3] 松尾,白井,発音情報を用いた訳語対の自動抽出，情報処理学会研究報告, NL-116-15, (1996)

[4] 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)ー記事の特徴分析と文の対応関係の検討ー，電気関係学会九州支部第48回連合大会(1995)

[5] 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)

表 6　出現頻度の多い名詞項目リスト

日付	１日	２日	３日	４日
英文から切り出した単語	Bank of Japan 日銀 (4)	Bank of Japan 日銀 (5)	Sakura Bank さくら銀行，さくら銀 (2)	Hitachi 日立 (4)
	Asian アジア (5)	Osaka 大阪府，大阪市 (5)	Sega Enterprises Ltd. セガ・エンタープライゼス (2)	Ministry of Finance 大蔵省 (5)
	Tokyo 東京，東京都 (14)	Tokyo 東京，東京都 (10)	Tokyo 東京，東京都 (2)	Tokyo 東京，東京都 (13)

日付	５日	６日	７日
英文から切り出した単語	Matsushita Electric Industrial Co. 松下電器産業，松下電工 (2)	Shoko Chukin Bank 商工中金 (1)	Toshiba Corp. 東芝 (3)
	Tokai Bank 東海銀行，東海銀 (2)	Small Business Finance Corp. 中小企業金融公庫 (1)	NEC Corp. ＮＥＣ (5)
	APEC ＡＰＥＣ (3)	Tokyo 東京，東京都 (1)	Tokyo 東京，東京都 (6)

日付	８日	９日	１０日
英文から切り出した単語	Kanagawa Prefecture 神奈川県 (6)	Tokyo Stock Exchange 東京証券取引所，東証 (4)	Mitsubishi 三菱電機 (6)
	Osaka 大阪府，大阪市 (6)	Hitachi Ltd. 日立製作所，日立 (5)	Osaka 大阪府，大阪市 (6)
	Tokyo 東京，東京都 (10)	Tokyo 東京，東京都 (13)	Tokyo 東京，東京都 (14)

英単語列
日本語対訳(出現数)