新聞記事における日英対応コーパスの自動構築

高橋大和†，松尾義博‡，古瀬蔵†
†NTTサイバーソリューション研究所
‡NTTコミュニケーション科学基礎研究所

1.はじめに

近年、インターネットの拡大と普及にともなって、様々な分野の電子化文書が公開され始めてきており、自然言語処理に有用な電子化文書の入手が容易になってきている。しかし、電子化文書の多くは単一の言語でのみ書かれており、日本語と英語それぞれで書かれているような翻訳文書などは数も少なく、利用できる文書量も限られている。インターネットで公開されている文書に関しては、著作権や文章表現の多様さも問題となる。現在の自然言語処理では、くだけた表現や省略の多い文章を扱うことは困難であり、分野限定された整った文体の文書が必要である。

筆者らは電話回線経由のパソコン通信有料情報サービスである日本経済新聞社のテレコンDBの新聞記事を、分野の限定された電子化文書として着目し、これをもとに大量の対訳コーパスを構築する研究を行っている。選んだ文書は、日本経済新聞社発行の四紙(日本経済新聞，日経産業新聞，日経流通新聞，日経金融新聞)の日本語記事とこれらの速報訳である Nikkei Telecom Japan News & Retrieval の英語記事である。これらの記事をもとに記事対応付けを行い、対訳コーパスの構築を目指す。文献[高橋 97]では、数値と固有名詞を利用して記事対応付けを行ってきた。この方法により、英語記事の約4割を正解率 100%で記事対応付けを行うことができたが、この方法に必要な固有名詞の英日対訳辞書を人手で作成していた。

本稿では、大量の対訳コーパスを自動的に構築できるよう、文献[松尾 96]の単語読み推定による訳語抽出方法を文献[高橋 97]の記事対応付けに適応し、英日対訳辞書の作成と記事対応付けを自動的に行う方法を提案する。また、記事対応コーパスから、数値と固有名詞による対応付けに統計的手法を組み合わせて日英文対応コーパスを構築する方法も提案する。

2.記事の対応付け

記事対応付けアルゴリズムを図1に示す。記事対応コーパスの構築は、記事単位ではなく、数日分をまとめて一括で行う。

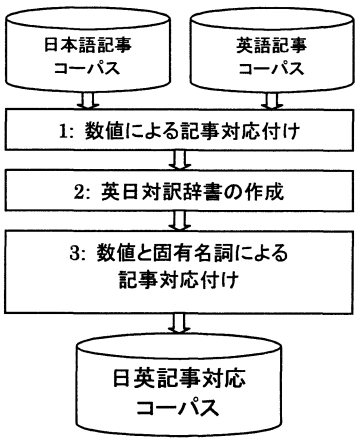

図 1 記事対応付けアルゴリズム

まず始めに、数値による記事対応付けを行い、大まかに記事対応コーパスを得る。数値による記事対応付けは、正解率も高く多くの記事対応を得ることができる[高橋 97]。この記事対応コーパスをもとに、文献[松尾 96]の単語読み推定による訳語抽出を行い、英日対訳辞書を作成し、これを利用して数値と固有名詞による対応付けを行って、より多くの記事対応コーパスを得る。

対応付けの方法は、日本語記事と英語記事それぞれからキーワードを切り出し、同じキーワードを多く含む記事を対応候補記事として選び、キーワードの個数で正しい対応かどうかを判定する[高橋 97]。

本節では、文献[高橋 97]の方法に新しく導入する単語読み推定による英日対訳辞書の作成について説明を行い、新しい記事対応付け方法の評価を行う。

2.1.英日対訳辞書の作成

表 1に、人手で作成した英日対訳辞書の項目の種別を示す。

表 1 英日対訳辞書の項目種別

種別	項目数
企業名	596
製品名	107
人名	23
地名・その他	136
合計	862

もっとも多い固有名詞は企業名である。日本の企業名は、「富士銀行」→"Fuji Bank"のように、固有名詞部分は読みをローマ字表記にすることが多く、外国の企業名は、"Motorola"→「モトローラ」のように、単語の綴りからカタカナ表記を推定できるものが多い。文献[松尾 96]の方法では、再現率 62%、適合率 98%の精度で自動抽出できることが示されており、この方法を組み込むことで、英日対訳辞書の自動作成を行う。

英日対訳辞書は、数値による記事対応付けで得られた対応記事ごとに、(1) 英語記事から固有名詞英単語列の切り出し、(2) 訳語対の抽出、を自動的に行い、その結果を英日対訳辞書に登録する。

2.1.1.固有名詞英単語列の切り出し

次のいずれかの条件を満たす単語列を固有名詞と考え、切り出しを行う。

・大文字を含む単語列

例 1: NTT Communication Science Lab.

例 2: SL-enhanced Intel i486SX

・大文字を含む単語の所有格に大文字を含む単語が連続する場合

例: International Standardization Organization's ISO9001

・大文字を含む単語の間に "of", "&"を挟む場合

例: Mitsubishi Trust & Banking Corp.

上の条件を満たす単語列に対して、以下の例外処理を行って、最終的に固有名詞として切り出す単語列を決定する。

・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は所有格の単語まで、また、"of" の後ろに大文字を含む単語が連接しない時は、"of"の前まで切り出す。

例: NTT's line → NTT

・ "The"、"A" は単語列に含まない。

例: The U.S. → U.S.

2.1.2.発音情報を用いた訳語対の抽出

2.1.1節の条件で英語記事から切り出した英単語列に関して、対応する日本語記事から訳語対の抽出を行う。

図 2 に訳語対の抽出アルゴリズムを示す。訳語対は、単語列の要素毎に抽出を行い、最終的に連結することで複数単語からなる長単位の訳語対を得る。

1. 日本語記事から切り出したカタカナ語と英語記事から切り出した単語列から訳語対を推定する。カタカナはそのままローマ字表記へ、英単語列は綴りから発音を推定してローマ字表記へ変換し、子音の並びと先頭の母音の比較を行い、その一致度で単語対応の判定を行う。
2. 1. で、英単語列から変換したローマ字表記からひらがなへ変換、日本語記事に対して形態素解析を行い、固有名詞もしくは未知語に 1 文字ごとに読みを付与してひらがなへ変換し、その一致度で単語対応の判定を行う。
3. 英単語列に対して、1.と 2.でとれる単語対応に企業名対訳辞書を併用して、対訳の組み合せを作成し、切り出した日本語固有名詞に同一の日本語訳が存在するかどうかで、長単位の訳語対を抽出する。

図 2 訳語対の抽出アルゴリズム

一ヶ月分の記事(1995 年 8 月分)から、この手続きにしたがって訳語対の抽出を行った実験結果を表 2に示す。

表 2より、編集距離が 0 である単語は訳語対として使えると考えられる。

ただし、失敗例として、漢字の場合、漢字の読みとして可能な組み合わせを網羅的に作り出すため、例えば、生地 = Shoji(商事)のように訳語対で無い単語にも一致してしまう。

また、編集距離が 1 の場合も正解率が比較的よいが、これは、正しい訳語対、アメリカ(の) = American のように、派生した言葉に対応付けられた訳語対、誤って対応付けられた訳語対などが混在しているので、場合分けをして評価を行う必要がある。

表 2 訳語対の抽出

編集距離 (Editing distance)	抽出数	正解数
0	852	825(96.8%)
1	287	247(86.1%)
2	301	158(52.5%)

本稿では、編集距離が 0 の訳語対を英日対訳辞書に登録する。また、最長の固有名詞を得るため

に、企業名によく使われる単語の対訳を登録した企業名対訳辞書(表 3)を用いて、「大和銀行」="Daiwa Bank"といった長単位の訳語対を得る。

表 3 企業名対訳辞書の例

英単語	訳語
Air Lines	航空
Bank	銀行
Express	運輸、通運
Foods	食品

自動作成した英日対訳辞書の評価を表 4に示す。

表 4 自動作成した英日対訳辞書

対象とした記事の期間	抽出数	正解数
1994/11/02-08	207	201(97.1%)
1995/08/01-31	878	837(95.3%)

1週間分の記事から自動作成した対訳辞書は、人手作成の辞書と共通の項目が 96 項目、新規に105 項目(うち、長単位は 18)を得た。人手作成の辞書は長単位の項目を優先して登録したため、共通項目が少なかった。また、新規に抽出された項目には、国名・地名(24)や一般名詞(15)もあり、これも人手作成辞書では登録していなかった。

2.2.評価実験

1週間分の記事(1994年11月2〜8日)を使って、記事対応付けの評価実験を行った結果を表 5に示す。

表 5 記事対応付け

英語記事総数：577 記事

方法	記事対応数	正解数
数値のみ	172	171 (99.4%)
数値＋ 1週間分の記事から 自動で作成した対訳辞書	186	185 (99.5%)
数値＋ 5週間分の記事から 自動で作成した対訳辞書	194	193 (99.5%)
数値＋ 1週間分の記事から 人手で作成した対訳辞書	214	213 (99.5%)

表 5より、1週間分の記事から作成した対訳辞書を使うことで数値のみの対応付けと比較して14 記事、1995 年 8 月分で抽出した英日対訳辞書を加え、5週間分(934 項目)の対訳辞書を用いることで、さらに 8 記事の正しい記事対応付けが行えることを確認した。これより、英日対訳辞書を拡張することで、対応記事を増やすことができる。

しかし、作成した辞書は短単語の項目が多く、現在の長単位を優先した英単語列の切り出しでは利用されない項目が多い。長単位の訳語抽出と利用法を改良することで、より多くの単語の対応が取れ、対応付けられる記事の数を向上できると考えられる。

また、対応付けに失敗した記事の内容を見ると、3,4,6%といった1桁の数値が、かなり長い対談形式の文章とマッチングしてしまったためであった。日英の記事の分量にかなりの差がある(英文5文に対して日本文 70 文)ため、対応付けの評価項目として、記事の分量比を考慮した対応付け記事の判定法の改良により、正解率を向上できると考えられる。

3.文の対応付け

現在、2節で構築した記事対応コーパスに対し、さらに文対応付けを行う方法の研究を行っている。対象としている日本経済新聞記事に関して、英語記事のほとんどの英文は対応する日本文があること[白井 95]から、1994 年 11 月 2〜8 日の記事対応済み 213 記事に対して、どの程度、内容の対応がとれるかの調査を行った。結果を表 6に示す。

表 6 英日記事間でとれる文対応

日付 / 評価	2日	3日	4日	5日	6日	7日	8日	合計
A	5	3	6	1	0	2	8	25 (11.7%)
B	32	15	34	12	3	40	46	182 (85.4%)
C	1	0	1	0	0	3	1	6 (2.8%)

評価は以下のように定めている。

A) 英語記事内容が日本語記事に含まれており、両方の情報量が等しい。
B) 英語記事内容が日本語記事に含まれており、日本語記事が英語記事より情報量が多い。
C) 英語記事内容の一部のみ日本語記事に含まれている。

表 6から、記事の 97.1%は英語記事の内容を日本語記事が含んでおり、文対応付けが行えると考えられる。評価 B が多いことから、日本語文には英語文に翻訳されていない文が多いであろうこ

とが予想できる。よって、文対応付けを行う場合は、基本的に英語文の情報を手掛りに、対応する日本語文を見つける方法がよいと考えられる。これを踏まえ、次のような文対応付けアルゴリズムを提案する。

3.1.文対応付けアルゴリズムの構成

文対応付けアルゴリズムを図3に示す。

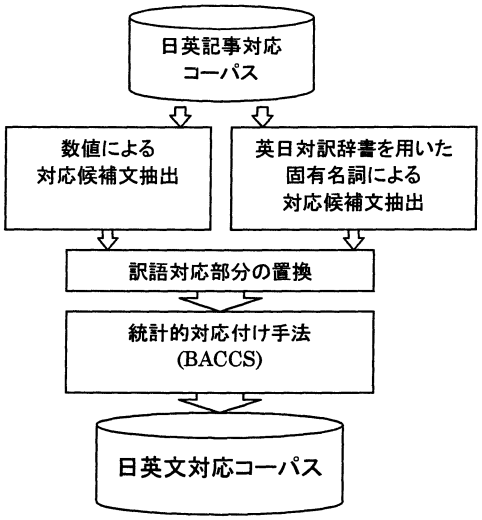

図 3 文対応付けアルゴリズム

2節で得られた日英記事対応コーパスに対して、数値と固有名詞それぞれを使って、英文に対応する候補文を抽出し、対応するキーワードを英語文・日本語文ともにマーク記号に置換しておく。置換後の記事に対して、文献[Haruno96]のBACCSを使い、対訳単語辞書と共起確率による文対応付けを行う。このとき、マーク記号は対応済み単語として扱われる。得られた対応候補文を単語対応の度合いで正しい対応付けかどうかを判定し、文対応コーパスとして出力する。

3.2.文対応付けの問題

現在、プロトタイプを作成し、日英記事対応コーパス1週間分を使って文対応付けの評価を行っている。数値による候補文抽出、固有名詞による候補文抽出処理は、記事対応付けで扱ったキーワードを利用しているが、より正確な文対応付けを行う上で、次のような課題がある。

[数値]
- 記事対応付けでは扱わなかった、年月日、曜日、時刻、株の銘柄番号を利用するための拡張が必要になる。
- 助数詞の拡張
  例: 「部屋」="rooms"
- "one" に関しては、代名詞との区別が必要

[固有名詞]
- 略称
  日本語記事で略称でも、英語記事で略称とは限らない。
  例: 日本経済新聞 → 日経
  例: Nippon Telegraph and Telephone → NTT

また、一番大きな問題となるのが、英語文1文が複数の日本語文と対応する場合をどう扱うかである。英語文が複数の節に切り分けられる場合であれば、節単位での対応付けも考えられるが、対応付けの評価をどう行うかという課題もある。

4.まとめ

大量の対訳コーパスを自動構築する方法として、単語読み推定による自動英日対訳辞書作成処理を組み込んだ記事対応付け方法の提案を行った。1994年11月2〜8日の記事1週間分を対象に評価実験を行い、英語記事の32.1%に対して対応する日本語記事を99.5%の精度で得ることができることを確認した。また、日英記事対応コーパスから、数値と固有名詞による対応付けに統計的手法を組み合わせて日英文対応コーパスを構築する方法の提案を行った。

今後は、記事対応付けシステムでは対応付けの評価の改良、文対応付けシステムでは方式の改良を効果の検証をしながら進めていく予定である。

参考文献

[高橋 97] 高橋,白井,大山,渡辺,上田:日英新聞記事の記事対応コーパス自動作成,言語処理学会第3回年次大会(1997)

[松尾96] 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, 96-NL-116-15 (1996)

[白井 95] 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)

[Haruno96] Haruno, M. and Yamazaki, T. High-Performance Bilingual Text Alignment Using Stastical and Dictionary Infomation, Proc. of the 34th Annual Meeting of ACL(ACL96)