分野固有の情報を利用した日英対訳記事コーパスの構築

5Y−4

松本 賢司　柏岡 秀紀　田中 英輝
(株)エイ・ティ・アール音声言語通信研究所

1. はじめに

著者らはコーパスに基づいた日英機械翻訳システムの研究を行なっている。これには大規模な対訳コーパスが必須である。そこで日本経済新聞社より提供された日本語新聞記事データとその英訳データからなる対訳コーパスの整備にとりかかっている。このデータの日本語と英語記事間にはほぼ忠実な翻訳関係があるが対応情報はない。本稿はこれらの記事間の対応を自動的に取る手法について報告する。

同様の試みは[1][2]などによって報告されているが人手による作業が必要であった。これに対して提案手法は既存の会社名対訳辞書などを使うことにより完全自動で対訳関係を同定することができる。本手法を使った実験によって約97%の精度（適合率）で記事間の対訳関係がとれたことを示す。

2. 対象データの特徴

日本経済新聞社は発行する四紙（日経本紙、産業、金融、流通）記事の一部を英語に翻訳して主に電子メディアを通じて配信している。過去の四紙記事、英文翻訳記事はそれぞれ別個のデータベース(DB)に蓄積されていて記事間の対応付けはされていない。表1に過去5年の日英記事数を記す。

表1 日経四紙とその翻訳記事の数
（上段:日経四紙記事／下段: 英文翻訳記事）

	'97	'98	'99	'00	'01
本紙	148242	150417	148542	144990	73278
	16393	14085	14915	15538	8790
産業	67424	66838	64532	62295	35569
	9794	10400	8822	8227	4409
流通	18773	18645	17751	17339	9674
	204	226	102	78	7
金融	28243	27906	27855	31335	16118
	4971	4791	4700	5347	2793

「本紙」は朝夕刊の合計、'01'は7月22日付まで

翻訳は基本的に1つの日本文記事から1つの英文記事を作る形式で行われている。

Automatic Alignment of Japanese-English Newspaper Articles using the MT system and bilingual dictionary of Company names
Kenji Matsumoto, Hideki Kashioka, Hideki Tanaka
ATR Spoken Language Translation Research Laboratories

3. 対応付け手法

著者らは上記の日英記事のうち、まず産業・企業記事中心の日経産業新聞とその翻訳記事を対象に対応付けを試みた。日英記事間の記事数の違いを考えて英文記事から翻訳元の日本文記事を見つけ出す事にした。この際、日本文記事DB上の全記事を対象とするのは処理効率上問題がある。そこで英文翻訳記事より得られる情報によって、日本文記事DBの中から日本文記事候補を事前に抽出し、これを対象に対応付け処理を行なう2段階処理方式をとった。

3.1. 候補記事の抽出

対応付け候補となる日本文記事を抽出するための情報には掲載日付と会社名を用いた。

a) 掲載日付を用いた候補抽出

英文翻訳記事の配信日と記事末尾の紙名・曜日の情報により翻訳元記事の掲載日をほぼ特定できる(図1。網掛け部は「日経産業新聞火曜日付」を翻訳の意)。これを日本文記事候補抽出の情報とした。

... choose between Microsoft Corp.'s Windows 2000 Server, Linux and other systems.
(The Nikkei Industrial Daily Tuesday edition)

図1 紙名・掲載日の情報

b) 会社名を用いた候補抽出

処理対象の日経産業新聞では会社名が記事の中心的な話題となる事が多く、翻訳に際しても省略されることはない。そこで英語記事に含まれる会社名を1つでも含む日本文記事は対訳記事の可能性があると考え、日本文記事候補として抽出する。

これを実現するために約35000社の会社情報からなる「日経会社属性ファイル」[3]を用い表2に示す会社名対訳辞書を作成した。

表2 日英会社名対訳辞書

コード	日本社名	英語社名
1911	住友林業	Sumitomo Forestry Co., Ltd.
9434	日本テレコム	Japan Telecom Co., Ltd
—	セイコーエプソン	Seiko Epson Corp.

英文記事からの会社名取得は以下のように行なった。店頭・上場企業については、記事中で会社名に続いて株式コードが表記される(図2の'1911')。そこで株式コードで会社名対訳辞書を参照して記事中の会社を特定し日本社名を取得した。

Sumitomo Forestry Co. (1911) is broadening activities designed to help lumber processors a ..

図2 会社名・株式コードの表記

非上場企業など株式コードの表記が無い会社名の抽出は、記事中より大文字で始まる単語の連続を取り出し[2]、日英会社名対訳辞書とマッチングした。記事中の会社名表記と対訳辞書の表記は、同一会社でも微妙にずれる場合がある（表3）。記事から抽出した文字列と対訳辞書中の社名表記の最長共通部分文字列を取り出した上で Dice 係数によって類似度を計算した。係数値が一定以上の会社の日本社名を対訳辞書から取得した。

表3 記事－対応表間の表記のずれ

Dell Computer Corp.,	----- (記　事)
Dell Computer KK	----- (記　事)
Dell Computer K.K.	----- (対応表)
Mitsui Hi-tec Inc.	----- (記　事)
Mitsui High-tec, Inc.	----- (対応表)

3.2. 対応付け処理

英文翻訳記事を市販の機械翻訳装置を用いて日本文に翻訳する。これと対応付け対象の日本文記事の形態素解析[4]結果から名詞を取り出して Dice 係数を用いた類似度を算出した。係数値の最大となった日英記事対を対訳関係にある記事対と判断した。

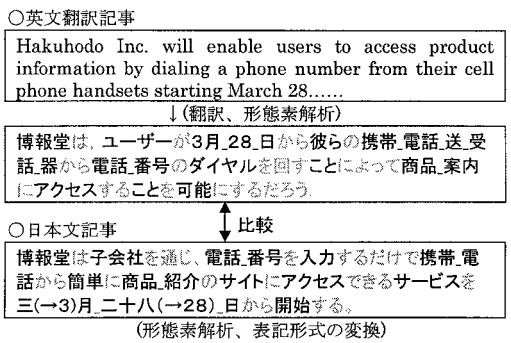

図3 対応付け判定処理

4. 対応付け実験

日経産業新聞を翻訳した3日分の英文記事 98 件（配信日 2001 年 3 月 7 日～3 月 9 日）を対象に翻訳元記事を特定する実験を行なった。

実験の結果を表 4 に示す。98 件中、84 件については掲載日付・会社名を使った候補記事抽出に成功した(A)。14 件については掲載日付のみによる候補抽出となった(B)。(A)記事中の 83 件は正しく対応付けられた。失敗した 1 件は誤った会社名のみを抽出したため候補記事の抽出に失敗していた。仮に対象掲載日付の全記事を対象に対応付け処理を行なっていれば正しい翻訳元記事を判定していた。(B)記事中の正しい対応付けは 12 件であった。全体では 98 件中、95 件（96.9%）の記事について正しい翻訳元記事を特定できた。

表4 対応付け実験

日付	記事数	(A)「日·会」抽出		(B)「日」抽出	
		対応OK	対応NG	対応OK	対応NG
3/7	38	33	1	4	-
3/8	33	27	-	5	1
3/9	27	23	-	3	1
計	98	83 (84.7%)	1	12 (12.2%)	2

5. おわりに

本稿では日英新聞記事間の対応付け手法を報告した。この手法は既存の会社名対訳辞書などを利用することで自動的に動作し精度についても高い値を示した。具体的な数字は報告していないが、事前に候補記事を抽出することにより対応付けに要する時間が大幅に短くなることも確認している。今後は、今回対象としなかった新聞記事を対象にした対応付け手法を検討するとともに文レベルでの対応付けなどを実現したい。

参考文献

[1] 白井ほか ：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,1373,p.855(1995)

[2] 高橋ほか ：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 回年次大会発表論文集,127 - 130(1997)

[3] 日本経済新聞社 「日経会社属性ファイル」

[4] 日本語形態素解析システム「茶筌(ChaSen) version 2.2.6 」(2001)