Internal Use Only（非公開）

TR－SLT－0075

NE タグつき日英放送ニュース記事コーパスの構築と基礎調査

Construction and Analysis of Japanese-English Broadcast News Corpus with Named Entity Tags

熊野正　Tadashi Kumano
柏岡秀紀　Hideki Kashioka
田中英輝　Hideki Tanaka
福島孝博　Takahiro Fukusima

2005年3月24日

概要

我々は、直訳でない、content-aligned な文書対から、固有表現（NE）抽出技術を利用して対訳 NE 対を獲得することを目指している。本研究のために、我々は日英対訳ニュース原稿 2,000 記事対に対して NE タグを付与したタグつきコーパスを構築した。本コーパスには、日本語/英語の各文書中の NE の出現、そして日本語/英語の各文書内および日英文書間での NE 間の共参照情報が付与されている。コーパスを分析した結果、例え直訳でない文書対であっても、各言語文書に出現する NE の種類やその出現順序はかなり類似しており、この性質を用いることで対訳 NE 対を獲得できることが期待できる。

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目2番地2 TEL：0774－95－1301

Advanced Telecommunications Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288,Japan
Telephone:+81-774-95-1301
Fax :+81-774-95-1308

1 はじめに

日本語や英語などさまざまな言語で、固有表現 (named entity; NE) 抽出の研究が進展している。英語などの Message Understanding Conference (MUC)[1]、日本語の Information Retrieval and Extraction Exercise (IREX)[2] など、コンテスト形式のワークショップも行われ、その結果、英語では抽出性能は実用レベルの近くまで達し [12]、NE 境界の発見が英語より難しいとされる日本語でも比較的高い性能をあげている [14]。

従来の NE 抽出研究の多くは単言語に閉じている。しかし、対訳文書に対して NE 抽出技術を適用することで、例えば対訳 NE 対の獲得などの応用が期待できる。我々は、ニュースや時事問題に関する文書など、NE が多く含まれる文書の日英機械翻訳を目指している。このような文書の翻訳において、NE が正しく翻訳できることは正しく内容の伝達に不可欠である。しかし、NE の翻訳は一般に既存の対訳辞書には載っていないため、最新の対訳文書などから NE の翻訳知識を獲得することは有用である。

対訳コーパスから翻訳などの知識を獲得する場合、例えば複数の公用語で作成される公文書のような、逐語的に近い対訳文書を用いる方が、実現も容易で得られる情報も多い。しかし、そのようなコーパスは最新の NE を大量に収集する対象としては魅力に欠ける。また、そもそも我々が対象にしている日英の言語対にはそのようなコーパスは非常に少ない。そこで我々は、複数言語で書かれたニュース記事のように、直訳ではない対訳コーパスから対訳 NE 対を獲得することを検討する。このようなコーパスは、日々新しい NE を含む文書が追加されるため、収集の対象に適している。しかし、対訳文書対の各々はおおよそ同じ内容を表わしている (content-aligned) ものの、一般に文書より小さな単位での言語間対応は自明でなく、対訳対の獲得はより難しい。

対訳コーパスから対訳対を抽出する際には、文単位での対応づけ [2, 5, 7, 10, 16] を出発点とすることが多い。しかし、非直訳文書対に文対応を与えることは一般に難しい。また、従来の非直訳コーパスを扱う統計的手法 [4] では低頻度語句の対訳を発見することは難しく、NE のようにコーパス中に 1 度しか出てこないことも多いような語句の対訳対抽出に用いるには網羅性の点で問題がある。従って、対訳 NE 対の抽出に特化した手法が必要である。対訳 NE 対を扱った従来の研究として、対訳文書中の NE に対してその対訳を相手言語側から発見するのに翻字処理 (transliteration) を用いるものがある [1, 6, 15]。しかし、この手法は人名や地名などには有効だが、NE 一般に対して適用可能ではない。

そこで我々は、対訳ニュース記事のような対訳文書対が、同じ話題を伝達する目的で作成されたものであることに着目する。NE は文書の内容を表現する根幹であるため、直訳でない対訳文書間でも NE の出現はかなり保存されていることが予想できる。従って、対訳文書対の各々に対して (単言語の) NE 抽出技術を適用して NE を網羅的に発見しておけば、各 NE の種別や出現位置・順

[1] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

[2] http://nlp.cs.nyu.edu/irex/

序の類似性を手がかりにして、言語間で対訳 NE 対を対応づけられることが予想できる。

2 言語文書対に対する NE 抽出・対応づけ技術は、対訳 NE 対抽出以外にもさまざまな応用が考えられる。例えば、

- 直訳でない対訳文書間に対して文などの単位で対応づけを行うことは、一般に困難な問題である。もし NE の対応づけが与えられれば、そのような対応づけの手がかりになり得る。
- 任意の 2 言語文書があるとき、文書間の内容の類似度を NE 対応の良さで見積もることができるかもしれない。

我々は、NHK の日英ニュース原稿 [11] に人手で NE タグを付与したタグつきコーパスを構築し、現象の分析や対訳 NE 抽出実験を行うことにした。このコーパスに与えられる NE タグは、日本語 NE 抽出ワークショップ IREX NE タスクの仕様を英語にも拡張して採用した。さらに、NE 間の共参照情報を、単言語文書内と、対訳文書間の 2 種類（以後、それぞれ「言語内」「言語間」と呼ぶ）付与することで、対訳 NE 抽出研究に有用なものにする。

本稿では、まず 2 章で、対象とする対訳コーパスの特徴を紹介する。次に、対訳 NE 抽出研究のためにはどのような方針でタグを付与すればよいかを検討し、その結果決まったタグづけの仕様を説明する。そして、この仕様に基づいて行ったタグつきコーパスの構築作業を概説し、作業の結果判明した問題点を示す。3 章では、構築したタグつきコーパスを分析し、NE の出現傾向や訳され方などを調査した結果を示す。最後に 4 章で、対訳 NE 抽出へ向けた今後の課題について述べる。また、付録として、タグづけの作業指示書を収録する。

2 NE タグつき日英放送ニュース記事コーパスの構築

2.1 NHK 日英ニュースコーパスの特徴

我々が NE タグつき日英コーパスの構築に利用する NHK 日英放送ニュース原稿は、日本国内に放送される日本語ニュース原稿と、これを海外向けの放送や国内向け 2 か国語放送のために翻訳した英語ニュース原稿の対からなる。

図 1 に日英ニュース原稿対の例を示す。英訳原稿は元の日本語原稿と同じ事柄を伝達する内容であるが、内容の詳細には多くの違いがある。これらの相違の原因には、以下のようなものが考えられる [11]。

視聴者の違い 特に海外向け放送用の原稿の場合、視聴者の興味や背景知識の違いに応じて、内容の削除や説明の補足などが、時として大幅に行われる。

放送日時の違い 英訳原稿は元の日本語原稿から遅れて放送されることが多い。そのため、日時に関する表現が変化したり、その間に起こった出来事に合わせて内容が変更されたりすることがある。

日英のニューススタイルや言語の違い 日英のニュース記事を見比べると、事実の提示順序など、好まれるスタイルが異なることが観察できる。日本語と英語の言語的な違いや文化的な違い

元の日本語記事:

1: 地震が続いている伊豆諸島できょう午前六時四十二分頃強い地震があり式根島で震度五弱を観測しました。

2: このほか震度四が新島、神津島、震度三が利島、三宅島、また関東各地や静岡県の一部で震度二や一の揺れを観測しました。

3: この地震による津波の心配はありません。

4: 気象庁の観測によりますと震源地は新島・神津島の近海で震源の深さは十キロ、地震の規模を示すマグニチュードは五点一と推定されています。

5: 六月末から地震活動が始まった伊豆諸島では活動が活発な状態とやや落ち着いた状態を繰り返していて、先月三十日も三宅島で震度六弱の強い地震を一回観測した他震度五強の地震が二回起きました。

6: これらの地震を含めて一連の地震活動では神津島や新島、三宅島で震度六弱の強い揺れを四回観測したのを含めてこれまでに震度五弱以上の地震が十七回起きています。

翻訳された英語記事:

1: A strong earthquake jolted Shikine Island, one of the Izu islands south of Tokyo, early on Thursday morning.

2: The Meteorological Agency says the quake measured five-minus on the Japanese scale of seven.

3: The quake affected other islands nearby.

4: Seismic activity began in the area in late July, and 17 quakes of similar or stronger intensity have occurred.

5: Officials are warning of more similar or stronger earthquakes around Niijima and Kozu Islands.

6: Tokyo police say there have been no reports of damage from the latest quake.

図 1　NHK 日英ニュース原稿の例

がこのような違いに影響していると考える。

2.2　NE タグづけ方針

人手による NE タグづけの仕様を決定するにあたっては、対訳 NE 対抽出研究への有用性や作業の効率を考慮し、以下の方針を定めた。

- NE 認定の基準は、既存の NE タグつきコーパスとの整合性をとる。MUC や IREX の認定基準は多くの議論の結果決められたものであり、尊重に値する。また、既存のタグつきコーパスや、これらを前提とするシステムとの相互運用も可能になる。
- 「言語内」共参照情報と「言語間」共参照情報を付与する。逐語訳でない対訳文書では、文書内に同じ対象を指し示す NE が繰り返し出現しているとき、その一つ一つが対訳文書中のどの表現と直接的な翻訳関係にあるのかは一般に自明でない。そこで、各言語文書内で同じ対象を指し示す NE のグループ情報（「言語内」共参照情報）と、対訳文書間で同じ対象を指し示す NE グループ間の対応情報（「言語間」共参照情報）の 2 種類の情報を付与することとし、実際の個々の NE の翻訳関係は付与しない。対訳 NE 対抽出研究のためのデータとしては、この情報で十分である。

表 1　NE 種別と具体例 [13]

NE 種別	例（英語）
「狭義の」固有表現:	
ORGANIZATION	The Diet; IREX Committee
PERSON	(Mr.) Obuchi; Wakanohana
LOCATION	Japan; Tokyo; Mt. Fuji
ARTIFACT	Pentium Processor; Novel Prize
時間表現:	
DATE	September 2, 1999; Yesterday
TIME	11 PM; midnight
数量表現:	
MONEY	100 yen; $12,345
PERCENT	10%; a half

- 言語内共参照情報は、NE どうしの間にのみ付与する。本来 NE は代名詞やその他一般表現とも同じ対象を指し示すが、タグづけ作業の簡便化のため、そのような情報は付与しない。

2.3　タグづけ仕様

1. タグの書式、NE 種別、NE 認定基準とも、日本語 NE 抽出ワークショップ IREX NE タスクのタグ仕様 [9] に準拠する。
   IREX NE タスクで定義されている NE 種別は、MUC-7 でも採用されている 7 種類（「狭義の」固有表現 3 種類、時間表現 2 種類、数量表現 2 種類）に人工物（商品などの具体物、法律や著作物などの抽象物）の名前 ARTIFACT を加えた 8 種類である。NE 種別の一覧を表 1 に示す。
2. 英語側のタグづけにも IREX の NE 種別を用い、NE 認定基準をできる限り適用する。NE 中の前置詞や冠詞の扱いなど英語特有の判断基準が必要な場面では、MUC-7 NE タスク定義書 [3] も併せて参照する[3]。
3. IREX 仕様の SGML タグ書式を拡張し、各 NE タグに以下の 2 種類のタグ属性を追加することで、言語内共参照情報と言語間共参照情報を表現する。
   ID="group_ID"　（必須）
   各 NE に属性 ID を与える。その値 *group_ID* は、各言語の文書内で同じ NE 属性でかつ同じ具体的対象を指す NE が一意になるような、グループ番号である[4]。同じ表記の NE

[3] IREX NE タスクのタグ仕様も MUC の議論をふまえて作成されており、両者に大きな不整合はない。
[4] 異なる文書内の ID の間には一意性はない。

日本語:
地震が続いている<LOCATION ID="1" COR="2">伊豆諸島</LOCATION>で<DATE ID="2" COR="4">きょう</DATE><TIME ID="3" COR="5">午前六時四十二分</TIME>頃強い地震があり<LOCATION ID="4" COR="1">式根島</LOCATION>で震度五弱を観測しました。 ···

英語:
A strong earthquake jolted <LOCATION ID="1" COR="4">Shikine Island</LOCATION>, one of the <LOCATION ID="2" COR="1">Izu islands</LOCATION> south of <LOCATION ID="3">Tokyo</LOCATION>, early on <DATE ID="4" COR="2">Thursday</DATE> <TIME ID="5" COR="3">morning</TIME>. ···

図 2 対訳 NE タグの付与例

同士だけでなく、例えば姓名と姓のみ、正式名称と略称のような異表記 NE 同士も同じ ID 番号を与える。原則として、文書内で、同一 NE 種別でかつ同一表記を持つ NE は全て同じ対象を指すと見なし、同じ ID 番号を与える[5]。

COR="cor_group_ID" （任意）

ある NE グループと同じ NE 種別でかつ同じ具体的対象を指す NE グループが対応する相手言語側にもあるときには、対応する両言語の NE グループの各要素に COR 属性を与える。その属性値 *cor_group_ID* は、対応する相手言語側 NE グループの ID 番号とする。

本仕様によるタグづけ例を図 2 に示す。

2.4 タグつきコーパスの構築

1995 年 3 月から 2000 年 4 月までの NHK 日英ニュース原稿の中から無作為に選んだ 2,000 記事対に対して、タグ付与作業を完了した。作業は、翻訳と言語データ作成の業務経験のある作業者によって行われた。

2.5 問題点

タグづけ仕様決定の議論や実際のタグづけ作業の過程で、現在のタグづけ仕様の持つ問題点がいくつか明らかになった。これらは、タグづけ作業を混乱させたり、作業成果の有用度を低下させうる。

代表的な問題と、現時点での対処を以下に示す。

[5] 例外あり。2.5.3 節参照。

2.5.1 日英のタグづけ粒度の相違

IREX の仕様より、日本語の NE は、形態素よりも小さな単位でも認定する。一方英語の NE は、MUC-7 の仕様にのっとり、単語より小さな単位のものは認定しない。この日英の粒度差が原因で、日本語側で認定できた NE に対して、英語側に対応表現があるにもかかわらず、対応する英語側 NE を認定できないことがある。現時点では、このようなものに対しては対応を付与していない。

日	<LOCATION>日本</LOCATION>人
英	Japanese

2.5.2 非直訳性

NE の翻訳にも、一般の表現のような非直訳性の問題がある。

- 意味的に対応関係にある日英表現間で、NE 種別が異なったり片方が NE でなかったりするために対応関係を付与できないことがある。例えば、日本語記事中で、日本の政府のことを指して「政府」という NE でない表現が使われたとき、これが英語側では “Japan” という NE に訳されることがしばしばある。このような場合には、英語側のみを NE に認定し、対応は付与しない。
- NE が意訳されてしまっているために、文書内の情報だけでは対応の認定が困難なことがある。タグづけ作業にあたっては、作業者に対し、できるだけこのような対応づけを調査して認定するよう依頼した。
  - 日本語記事中の相対的な時間表現は英語では絶対表現に訳されることが多い。対応は元記事が書かれた日時が分からないと認定できない。
  - 日本語記事中の金額表現は英語側ではドルに換算されて表現されることが多い。対応は元記事が書かれた時点の通貨レートが分からないと認定できない。例えば、日本語記事中の「三千億円」が、当時の円–ドルレートに基いて概算した結果、英語側で “three billion U-S dollars” と翻訳されている例があった。

2.5.3 同一対象を指すかどうかの判断

タグづけ仕様の中で、同一 NE 種別で同一表記の NE は、個別の判断なしに同一対象を指す NE と見なし、同じ NE グループ ID を与えてよい、と定めた。これは、判断を簡便にするためである。しかし、特に時間表現や数量表現において、この基準を適用するのが困難なケースが考えられる。

例えば、対訳表現「先週の日曜と今週の日曜」– “last Sunday and this Sunday” を考えてみる。日本語側で IREX の基準に従って NE を認定すると、NE 内に助詞「の」を含むことができないため、時間表現「先週の日曜」「今週の日曜」のどちらからも「日曜」のみが認定される。これに対して、英語側は、MUC-7 の基準より、“last Sunday” と “this Sunday” のそれぞれ全体が認定される。

ここで、日本語側の 2 つの「日曜」を同じ NE グループにしてしまうと、英語側の NE との対応

表2　コーパスの文数・単語/形態素数

	記事数	文数 （記事あたり）	単語/形態素数 （記事あたり）
日	2,000	10,803 (5.40)	594,832 (293.4)
英		13,950 (6.98)	318,949 (159.5)

づけがうまくいかなくなる。そこで現時点では暫定的に、このような問題が発生したときにのみ、同一 NE 種別で同一表記の NE が異なるグループに属することを認め、対応づけを行うようにした。この例の場合は、以下のようになる。

日	先週の　<DATE ID="1">日曜</DATE>　と 今週の　<DATE ID="2">日曜</DATE>
英	<DATE COR="1">last Sunday</DATE>　and　<DATE COR="2">this Sunday</DATE>

日英の NE 認定基準の整合性については、再検討をする必要がありそうだ。

3　分析

対訳 NE 対自動獲得の可能性を探るために、今回タグづけを行った日英ニュース原稿 2,000 記事対に対して、対訳 NE 抽出の実現に必要な基礎調査を行った。

3.1　コーパスの規模

コーパスの文数と単語/形態素数を表 2 に表す。日本語側の形態素数と英語側の単語数は単純に比較できるものではない。しかし我々は、同じ内容の日英の文章に含まれる日本語の形態素数と英語の単語数は比較的似通っているという経験を持っており、それを考え合わせると、英語側の単語数は有意に少ないと言える。従って、翻訳によって記事の内容は減少する傾向にあると言うことができる。

3.2　NE 出現の単言語特性

3.2.1　出現数

コーパス中に出現した NE の、NE 種別ごとの出現数/グループ数一覧を表 3 に示す。表中、「グループ」数とは、各記事に含まれる NE グループの総数であり、言い換えると、各記事中に出現した「異なる具体的対象」の総数である[6]。

NE 種別ごとの出現数の比率は、MUC-7 や IREX で用意された英語や日本語のデータと同様の

[6] 異なる記事中に出現する同一の具体的対象を指し示すグループを同一視して数えた、「具体的対象の種類」の数ではない。

表 3　NE 出現数

	日本語				英語			
	出現	記事あたり	グループ	記事あたり	出現	記事あたり	グループ	記事あたり
合計	43,420	(21.71)	22,944	(11.47)	28,341	(14.17)	18,630	(9.32)
ORGANIZATION	9,726	(4.86)	4,784	(2.39)	5,343	(2.67)	3,446	(1.72)
PERSON	6,050	(3.03)	2,855	(1.43)	4,782	(2.39)	2,426	(1.21)
LOCATION	15,779	(7.89)	6,811	(3.41)	10,402	(5.20)	6,043	(3.02)
ARTIFACT	703	(0.35)	406	(0.20)	348	(0.17)	268	(0.13)
DATE	7,764	(3.88)	5,273	(2.64)	5,357	(2.68)	4,702	(2.35)
TIME	1,459	(0.73)	1,227	(0.61)	454	(0.23)	419	(0.21)
MONEY	1,018	(0.51)	820	(0.41)	964	(0.48)	675	(0.34)
PERCENT	921	(0.46)	768	(0.38)	691	(0.35)	651	(0.33)

傾向を示している。NE の出現に関して、これら既存のデータと同様の性質を持ったデータであると言える。

日本語から英語への翻訳による NE 出現数の変化を見てみると、翻訳の結果、出現 NE 数は減少する傾向にある。しかし、全単語/形態素数の減少の度合（3.1 節参照）ほどではなく、特に NE 異なり数は日本語に近い数になっている。翻訳によって比較的 NE は保存されていることがわかる。ちなみに、NE 種別 TIME の減少のみ著しいが、これは、英訳されたニュースは海外向けなどに日をおいて放送される場合が多く、時刻情報はあまり重要ではないと考えられているためではないかと思われる。

3.2.2　NE グループ内の NE の特性

同一グループ ID を持つ NE が記事中にどの程度繰り返し現れるか、それらの表記がどの程度揺れているかを調査するために、コーパス中各記事中で同一グループ ID を持つ NE の集合（NE グループ）の要素数（*freq*）、各 NE group 中の異表記数（*sort*）、そして、次式で計算される、同一グループ ID の 2 つの NE が異なる表記を持つ割合（*uniq*）を、それぞれ調査した。

$$uniq = \frac{{}_{freq-2}C_{sort-2}}{{}_{freq-1}C_{sort-1}} = \frac{sort-1}{freq-1} \quad (freq \geq 2)$$

これらの NE 種別ごとの平均値を表 4 に示す。

表 4 から、全般的には、英語の方が同一グループ ID の NE 出現数 *freq* は少ない一方で NE 表記の種類 *sort* は多く、結果として英語の方が *uniq* が大きくなることが分かる。これは、英語では習慣的に同一表現の繰返しを嫌って代名詞化や言い替えを行うが、日本語ではそのような習慣はないことに起因すると思われる。この性質は、以下に示すように、NE 種別によって若干異なる。

- 英語の PERSON の *sort* が顕著に多い。これは、英語では人名を表現するのにまず姓名（フルネーム）を示し、2 回目以降は姓のみを示すことが多いためである。日本語側では、とり

表4　NEグループの大きさと内部の表記揺れ

	日本語			英語		
	freq	*sort*	*uniq*	*freq*	*sort*	*uniq*
平均	1.89	1.11	0.144	1.52	1.14	0.332
ORGANIZATION	2.03	1.13	0.151	1.55	1.17	0.358
PERSON	2.12	1.13	0.134	1.97	1.48	0.641
LOCATION	2.32	1.14	0.111	1.72	1.06	0.098
ARTIFACT	1.73	1.08	0.119	1.30	1.05	0.181
DATE	1.47	1.09	0.200	1.14	1.03	0.218
TIME	1.19	1.04	0.229	1.08	1.01	0.172
MONEY	1.24	1.04	0.148	1.43	1.40	0.958
PERCENT	1.20	1.03	0.151	1.06	1.02	0.414

わけ有名人の場合などは、始めから姓のみを示すことがほとんどである。

- 日本語のLOCATIONの*sort*が英語より多い。これは、日本語では組織名を1度しか示さない時にも、その略称や頭字語が一般的な呼称であるならば、正式名称と併せて示すことが多いからである（例えば、「…ＯＰＥＣ、石油輸出国機構…」）。
- 英語のMONEYの*sort*が非常に多い。これは、翻訳先である英語側では、日本語側で元々使われていた通貨単位（円など）での金額と、ドルなどに換算した金額を併用することが多いからである。
- 時間表現や数量表現の*freq*は、狭義の固有表現に比べて少ない。

3.3　NEの翻訳特性

3.3.1　対訳存在率

あるNEの出現に対応する相手言語側のNEが存在する確率（出現対訳存在率）と、あるNEグループに対応する相手言語側のNEグループが存在する確率（グループ対訳存在率）を表5に示す。英語側に存在しているNEには対応する日本語NEが高い確率で存在していることが分かる。

また、表4と同様のNEグループの要素数や表記の揺れに関する調査を、対応する相手言語側NEが存在の有無で分類して再度行った結果を、表6に示す。両者を比較すると、対応する相手言語側NEが存在するものの方が、*freq*が顕著に大きいことが分かる。これは、繰り返し言及されるNEの方が重要な内容を表していて翻訳されやすいことの表われであると考える。この性質は対訳NE対の発見の重要な手がかりとなり得る。

3.3.2　NE出現順序の保存率

記事中でのNEの出現順序が相手言語側でどの程度保存されているかを調査した。content-alignedな文書対の性質上、言語間に個々のNEの出現間の対応関係は特定できないため、順序の

表 5　NE 対訳存在率

	日 → 英		日 ← 英	
	出現	グループ	出現	グループ
平均	0.747	0.649	0.851	0.799
ORGANIZATION	0.695	0.613	0.888	0.851
PERSON	0.877	0.775	0.947	0.913
LOCATION	0.792	0.674	0.833	0.760
ARTIFACT	0.681	0.599	0.925	0.907
DATE	0.736	0.679	0.780	0.761
TIME	0.515	0.500	0.645	0.637
MONEY	0.648	0.638	0.820	0.775
PERCENT	0.730	0.711	0.844	0.839

表 6　対応先の有無と、NE グループの大きさ・内部の表記揺れ

	対応先	日本語			英語		
		freq	*sort*	*uniq*	*freq*	*sort*	*uniq*
平均	有	2.18	1.15	0.147	1.62	1.17	0.341
	無	1.36	1.04	0.134	1.13	1.03	0.216
ORGANIZATION	有	2.30	1.18	0.164	1.62	1.19	0.365
	無	1.60	1.05	0.113	1.16	1.04	0.250
PERSON	有	2.40	1.15	0.122	2.05	1.52	0.636
	無	1.16	1.05	0.334	1.20	1.15	0.824
LOCATION	有	2.72	1.18	0.115	1.89	1.08	0.104
	無	1.48	1.04	0.094	1.20	1.01	0.049
ARTIFACT	有	1.97	1.11	0.139	1.33	1.05	0.184
	無	1.37	1.02	0.060	1.04	1.00	0.000
DATE	有	1.60	1.12	0.197	1.17	1.04	0.226
	無	1.21	1.04	0.214	1.05	1.01	0.146
TIME	有	1.39	1.11	0.292	1.10	1.02	0.217
	無	1.13	1.03	0.184	1.06	1.00	0.000
MONEY	有	1.26	1.05	0.178	1.51	1.48	0.954
	無	1.21	1.01	0.068	1.14	1.14	1.000
PERCENT	有	1.23	1.04	0.178	1.07	1.03	0.422
	無	1.12	1.00	0.023	1.03	1.01	0.333

保存性を議論することができない。そこで、以下のように、記事中に出現する 2 つの NE グループの出現順序を定義し、NE グループの順序保存性を調査することにした。

1. NE グループの要素のうち記事中で最初に出現するものの出現位置を、その NE グループの「初出位置」とする。
2. 記事中に出現する 2 つの NE グループの出現順序を、各 NE グループの「初出位置」の順序

表 7　NE 出現順序保存率

		任意の文間	同一文中を除く	
			日本語	英語
任意の種別間		0.775	0.846	0.817
同一種別間	平均	0.794	0.852	0.822
	ORGNIZATION	0.811	0.863	0.824
	PERSON	0.870	0.878	0.872
	LOCATION	0.752	0.834	0.800
	ARTIFACT	0.902	0.875	0.909
	DATE	0.809	0.852	0.815
	TIME	0.816	0.778	0.850
	MONEY	0.823	0.890	0.870
	PERCENT	0.905	0.909	0.907

と定義する。

NE グループの順序保存性は、相手言語側に対応する NE グループがある NE グループを一方言語の記事中から 2 つ選択したときに、それらの出現順序と各々の対応先 NE グループの相手言語側での出現順序とが一致する割合である、「順序保存率」で測定した。

まず最初に、任意の 2 つの NE グループを選択した場合の順序保存率を調査した。続いて、同一 NE 種別の NE グループを 2 つ選択したときの順序保存率について調査を行った。また、さらに、日本語側または英語側において記事中の初出位置が同一文中であるような組み合わせを除外した場合についても同様の調査を行った。調査結果の一覧を表 7 に表す。

調査結果から、対応関係のある NE グループの出現順序はかなりよく保存されているということができる。言い換えると、言及される具体的対象が記事中に登場する順番は両言語でかなり似通っている、ということである。またこの性質は、同じ種別の NE グループ間の順序関係において、より顕著である。さらに、一方の文書中で初出位置が異なる文内である 2 つの NE グループは、初出位置が同じ文内である組み合わせに比べて、相手言語側文書中で対応先の出現順序が保存されている可能性が高い。この理由の一つとして、調査対象である日本語–英語文書対においては、言語間の構文構造の違いによって、文内の順序の保存性がある程度損われているのではないかということが考えられる。

ここに挙げたような NE グループの出現順序に関する性質は、NE グループの言語間対応づけを推定する際に重要な手がかりとなり得る。

4　おわりに

本稿では、NE 翻訳知識獲得のための対訳 NE 抽出へ向けた取り組みとして、NE タグとその「言語内」および「言語間」共参照情報を付与した NE タグつき日英放送ニュース記事コーパスの構築

について報告した。また、コーパス中の 2,000 記事対を分析し、NE 種別や出現回数/出現順序などの情報が日英ニュース記事に現れる NE 間の対応推定の手がかりとして期待できる分析結果を示した。引き続き対応推定に必要な情報の検討を行うことで、高性能な NE 対応推定の実現を目指す。

参考文献

[1] Yaser Al-Onazian and Kevin Knight. Translating named entities using monolingual and bilingual resources. In *Proceedings of the 40st Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pp. 400–408, 2002.

[2] Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL-91)*, pp. 169–176, 1991.

[3] Nancy Chinchor. MUC-7 named entity task definition. `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html`, 1997.

[4] Pascale Fung and Lo Yuen Yee. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL '98)*, Vol. I, pp. 414–420, 1998.

[5] William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, Vol. 19, No. 1, pp. 75–102, 1993.

[6] Isao Goto, Noriyoshi Uratani, and Terumasa Ehara. Cross-language information retrieval of proper nouns using context information. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS 2001)*, pp. 571–578, 2001.

[7] Masahiko Haruno and Takefumi Yamazaki. High-performance bilingual text alignment using statistical and dictionary information. In *Proceeding of the 34th International Conference on Computational Linguistics (ACL '96)*, pp. 131–138, 1996.

[8] Lynette Hirschman and Nancy Chinchor. MUC-7 coreference named entity task definition. `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/co_task.html`, 1997.

[9] IREX 実行委員会. 固有表現抽出課題 (version 990214). IREX ワークショップ予稿集, pp. 265–273, 1999.

[10] Martin Kay and Martin Röscheisen. Text-translation alignment. *Computational Linguistics*, Vol. 19, No. 1, pp. 121–142, 1993.

[11] 熊野正, 後藤功雄, 田中英輝, 浦谷則好, 江原暉将. 翻訳用例提示システムの設計・開発・運用. 電子情報通信学会論文誌, Vol. J84-D-II, No. 6, pp. 1175–1184, 2001.

[12] Elaine Marsh and Dennis Perzanowski. MUC-7 evaluation of IE technology: Overview and results. `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/`

proceedings/muc_7_proceedings/marsh_slides.pdf, 1998.

[13] 関根聡, 井佐原均. IREX プロジェクト概要. IREX ワークショップ予稿集, pp. 1–5, 1999.

[14] Satoshi Sekine and Hitoshi Isahara. IREX: IR and IE evaluation project in Japanese. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)*, 2000.

[15] Bonnie Glover Stalls and Kevin Knight. Translating names and technical terms in Arabic text. In *Proceedings of the Workshop on Computational Approaches of the Semitic Languages*, pp. 34–41, 1998.

[16] Takehito Utsuro, Hiroshi Ikeda, Masaya Yamane, Yuji Matsumoto, and Makoto Nagao. Bilingual text matching using bilingual dictionary and statistics. In *Proceedings of the 32th International Conference on Computational Linguistics (ACL-94)*, pp. 1076–1082, 1994.

付録 A　日英対訳固有表現タグづけ作業の指示

A.1　作業指示書

固有表現タグづけ作業について (ver. 0.1)
2002/10/10
熊野 正 (tadashi.kumano@atr.co.jp)

1．作業セットの内容
hand-tagged_set.lzh を展開すると、以下の内容からなっている。

```
hand-tagged_set/ ... 日英記事対セット
  200001010129:200001010007/
    j-200001010129.txt ... 日本語記事
    e-200001010007.txt ... 英語記事
  200001010432:200001010012/
    j-200001010432.txt
    e-200001010012.txt
  ...
example/ ... 作業結果の例
  200004011346:200004010021/
    j-200004011346.txt
    e-200004010021.txt
  200004011380:200004020003/
    j-200004011380.txt
    e-200004020003.txt
  200004011384:200004020001/
    j-200004011384.txt
    e-200004020001.txt
```

2．作業手順

各日英記事対について：

1) 別紙「固有表現抽出課題 (version 990214)」の仕様に従い、
日本語、英語、それぞれ独立に固有表現タグを付与する。
固有表現タグを付与する範囲は、タイトル (###TITLE: 行) と本文である。
また、OPTIONAL タグ内の POSSIBILITY, TYPE 属性の属性値は、
POSSIBILITY="～", TYPE="～" のように " " で括ること。

英語記事にタグを付与する際には、以下の項目にも従うこと。
・表現前後の空白やカンマ、ピリオド、括弧類、引用符類はタグ中に含めない。
・表現の前の冠詞はタグ中に含めない。
ただし、それが固有表現を特に構成するものであるならば含めてもよい。
・表現の後の所有格表記 ('s や ') はタグ中に含めない。
ただし、それが固有表現を特に構成するものであるならば含めてもよい。
・単語の一部や acronym の一部をタグで括ってはならない。
ただし、'-' などで接合された合成語をその位置で分解し
各々にタグを付与することは可能である。
・(他にも随時追加予定)

2) 1) の結果タグが付与された日本語、英語記事の各々について、
タグを付与した固有表現に ID を付与する。
ID は、その記事内で
「同じタグ名を与えられた同じ表記の固有表現」は同じ番号を持ち、

それ以外は全て異なる番号を持つように付与する。

具体的には、
記事の先頭（タイトルも含んで）から最初にある

…<TAG_NAME>〜</TAG_NAME>…

という形式でタグが付与されている固有表現に対して

…<TAG_NAME ID="1">〜</TAG_NAME>…

というように ID 1 を付与する。
以後、順番に、出現するタグに対して ID 2, 3, … を付与していくが、
それ以前に出現したタグと同じタグ名で固有表現の表記も同じものに対しては、
新しい番号を付与せずに、
以前出現した「同じタグ名の同じ固有表現」の番号を与える。

3) 2) の結果タグに ID が付与された日本語、英語記事の各々について、
記事内の固有表現間の対応情報を付与する。
ここでいう「対応」とは、
同じタグ名を持ち、同じ対象を指し示すような内容の固有表現同士の関係をいう。
例えば、同じ人の名前がフルネームと姓だけの両方で表現されている場合や、
ある対象物の名前を正式名称と略称の両方で呼んでいる場合がそれにあたる。

具体的には、

…<TAG_NAME ID="a">〜〜</TAG_NAME>…

…<TAG_NAME ID="b">〜〜〜</TAG_NAME>…

の両者が同じ対象を指し示しているならば、

…<TAG_NAME ID="a" PAR="b">〜〜</TAG_NAME>…

…<TAG_NAME ID="b" PAR="a">〜〜〜</TAG_NAME>…

とする。
3 つ以上の固有表現の間に対応情報を付与するときには、
PAR="x,y" のように PAR 属性内部を , で区切って列挙する。
PAR 属性内部の ID の順序は任意でよい。

4) 3) の作業結果の日英記事対に対して、
日英の固有表現間の対応情報を付与する。
ここでいう「対応」とは、
同じタグ名を持ち、同じ対象を指し示すような内容の固有表現同士の関係をいう。
対応する固有表現が相手言語側に複数種類存在するときには、
最も直接的に対訳関係にあるものを特定できるなら、それを 1 つ指定する。
例えば、姓のみの人名に対し、相手言語側にフルネームと姓のみの両方が
存在するならば、姓のみのもの 1 つだけを対応の相手に指定する。
特定できない場合（どれも同等であるか、判断がつかない場合）には、
それらを全て対応の相手として列挙する。
具体的には、

日本語側：…<TAG_NAME ID="a">〜〜</TAG_NAME>…

英語側：...<TAG_NAME ID="b">?????</TAG_NAME>...

の両者が同じ対象を指し示しているならば、

日本語側：…<TAG_NAME ID="a" COR="b">〜〜</TAG_NAME>…

英語側: ...<TAG_NAME ID="b" COR="a">?????</TAG_NAME>...

とする。
相手言語側の固有表現 2 つ以上を列挙して対応情報を付与するときには、
COR="x,y" のように COR 属性内部を , で区切って列挙する。
COR 属性内部の ID の順序は任意でよい。

補足事項:

・タグ中に複数の属性 (ID, PAR, COL) を付与する場合、
それらの間の順序は任意である。

A.2 補足

1. 明らかなミスタイプの扱い
タグを付与される固有表現の中に明らかなミスタイプがあった場合に限り、
以下の方法で修正を行ってください。

例 1)
大洋神戸銀行はこの発表を…
↓
<ORGANIZATION>大<DEL ALT="陽">洋</DEL>神戸銀行</ORGANIZATION>はこの発表を…
例 2)
大陽陽神戸銀行はこの発表を…
↓
<ORGANIZATION>大陽<DEL>陽</DEL>神戸銀行</ORGANIZATION>はこの発表を…
または
<ORGANIZATION>大<DEL>陽</DEL>陽神戸銀行</ORGANIZATION>はこの発表を…
(どちらでもかまいません)

2. 英語の冠詞の扱い
英語の固有表現抽出コンテスト Message Understanding Conference (MUC-7)
Named Entity Task の定義書での扱いに準拠することにします。
具体的には、以下のとおりです。

a) 基本は、固有表現前後の冠詞は除外してタグづけする。
b) PERSON において、
それがある人 (々) の別称 (通称) であり、通常定冠詞と共に用いられる場合には、
定冠詞も含めてタグづけする。(from MUC-7 定義書 Appendix. A.1.6.2)
例) <PERSON>The Godfather</PERSON> ...
c) LOCATION において、
冠詞も地名の一部であるならば、冠詞も含めてタグづけする。
(from MUC-7 定義書 Appendix. A.4.4)
例) <LOCATION>The Hague</LOCATION>
d) ORGANIZATION において、
冠詞は通常タグ付けしない。(from MUC-7 定義書 Appendix. A.2.2.1)
e) ARTIFACT において、
冠詞も具体物名の一部であるならば、冠詞も含めてタグづけする。
例) "<ARTIFACT>The Fifth Element</ARTIFACT>" (映画名)

このメールに MUC-7 定義書を添付します。
冠詞の扱いや前置詞の扱いなど、参考にしていただきたいのですが、
あくまでタグづけの定義 (どのような固有表現にタグを与えるか、など) は
先にお渡しした IREX 定義書 (固有表現抽出課題 version 990214) に従います。

両者に矛盾があるときには IREX 定義書 の内容を優先します。

3．固有名詞（特に地名）の形容詞形や～人などの扱い
固有名詞（特に地名）の形容詞形や～人などにはタグを付与しません。
(from MUC-7 定義書 A.4.3)
例） American exporters

4．固有名詞の複数形
単数形の表記で固有名詞と認識されるものの複数形は、(e)s を除外してタグづけする。
例） They sold 2000 <ARTIFACT>Accrod</ARTIFACT>s last year.