二言語間の放送ニュース記事の自動対応付け

加藤　直人　　江原　暉将
ＮＨＫ放送技術研究所
E-mail: katonao@strl.nhk.or.jp, eharate@strl.nhk.or.jp

１　はじめに

機械翻訳の精度向上や翻訳メモリの充実には大規模な対訳コーパスが必要である．対訳コーパスを増やす１つの手段として，本来は対訳コーパスであるにもかかわらず記事対応付けが失われてしまった，単言語コーパスを利用することが考えられる．実際，このようなコーパスを対訳コーパスとするために，さまざまな記事（文書）の自動対応付けの手法が提案されている．これらの手法は，会社名や数量表現など原文の一部の情報のみを使って対応付けを行なっている[高橋 97][松本 01][Hasan 01]．しかし，ニュース記事のようにそれほど長くない文書の場合には，これらの情報だけでなく，記事中のさまざまな言語的情報を使うほうが自動対応付けの精度向上を期待できる．

本稿では，単語の出現順という情報も使い，記事の自動対応付けをする手法について述べる．本手法は機械翻訳を用いて翻訳し，翻訳結果と対訳候補を単語の出現順で比較する．この比較のために，単語出現位置辞書を導入している．

２　多言語の放送ニュース記事

２．１ニュース記事作成の流れ

ＮＨＫには電子化された２２言語の放送ニュース記事が存在する．これらの記事は，日本語記事を元にして，それぞれの言語に翻訳される．図１にその流れを示す．はじめにＴＶニュースなどに使われる日本語記事が作られる．それを元に二カ国語放送用などのために英語記事が作成される．残りの２０言語は短波やインターネットによるラジオ国際放送用に作成される．韓国語と中国語は日本語記事から作成され，それ以外の１８言語（フランス語～スワヒリ語）は英語記事から作成される．翻訳作業は毎日行われるので，各言語のコーパスは比較的容易に収集できる．しかし，翻訳先の記事が翻訳元のどの記事に対応するかを記述していない場合もあるため，対訳となっていない記事も少なくない．

２．２対応付けに用いるニュース記事の特徴

図２に対訳記事の一例として日本語記事とそれを人手で翻訳して得られた韓国語記事を示す．図２を見ると，数表現の１つである，日付「１２」や「１３」が一致しているのがわかる．このような数量表現は記事対応を付ける際の重要な手がかりとして有効であることが指摘されている[高橋 97]．しかしながら，放送ニュースは単純に翻訳するというわけではなく，その外国語ニュー

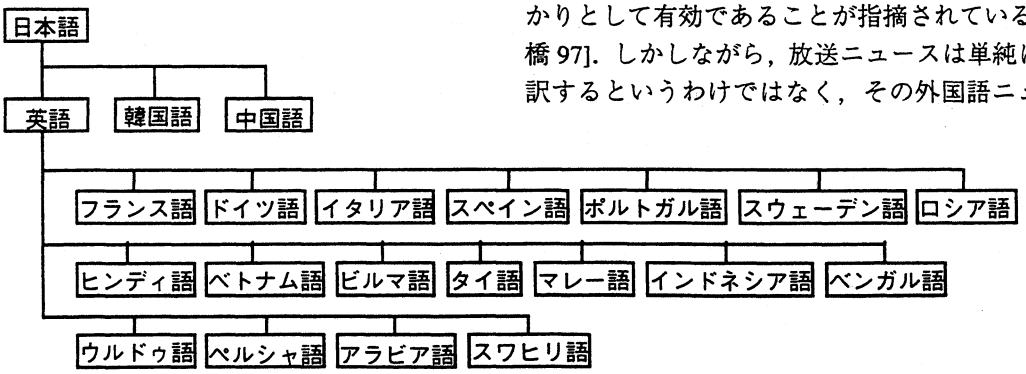

図１　翻訳の流れ

X:200106020637
N:カンボジアに１億ドル援助表明へ
D:2001年06月02日

J1：政府は、今月１２日から東京で開かれるカンボジアの支援国会合で、今年度にカンボジアに対しておよそ１億ドルの援助を実施し、復興を支援することを表明する方針です。
J2：カンボジアの支援国会合は、今月１２日と１３日の２日間、東京で開かれ、世界銀行や日本などの支援国がカンボジアの復興に向けた経済協力を話し合うことになっています。
J3：カンボジア政府は、かつての内戦で立ち遅れた経済基盤を整備するため、今後も３年間で１５億ドルの資金が必要になるとして各国に支援を要請していますが、日本政府は支援を行う前提として軍事費の削減や税金の徴収体制の強化によって財政状況を改善することや、ＯＤＡ・政府開発援助を巡る汚職を根絶するように求めることにしています。
J4：その上で、保健医療の改善や道路建設のための無償援助、カンボジアからの研修生の受入れなどの技術協力によって、今年度におよそ１億ドル・日本円にしておよそ１２０億円の援助を実施し、カンボジアの復興を支援することを表明する方針です。

(a) 日本語記事

D:2001/06/02

K1：캄보디아의 지원국회합이 이달 12 일과 13 일 이틀간에 걸쳐 도쿄에서 열려, 세계은행과 일본 등의 지원국들이 캄보디아의 부흥을 위한 경제협력을 논의할 예정입니다.
K2：캄보디아 정부는 내전으로 인해 뒤처진 경제기반을 정비하기 위해 앞으로 3 년간 15 억달러의 자금이 필요할 것으로 보고 각국에 지원을 요청하고 있습니다.
K3：일본정부는, 이번 회합에서 지원을 실시하는 전제로서, 군사비 삭감과 세금 징수체제의 강화로 재정상황을 개선할 것과 ODA/정부개발원조를 둘러싼 오직을 근절하도록 요구할 방침입니다.
K4：또, 보건의료의 개선과 도로건설을 위한 무상원조 등 금년도에 약 1 억 달러의 원조를 실시하고, 캄보디아의 부흥을 지원할 것을 표명할 방침입니다.

(b) 韓国語記事

図２　多言語記事の例（日本語と韓国語）

T1：カンボジアの 支援国会合が 今月 12日と 13日 二日間に わたり 東京で開かれて, 世界銀行と 日本などの 支援国らが カンボジアの 復興のため 経済協力を 議論する 予定です.
T2：カンボジア 政府は 内戦により 遅れた 経済基盤を 整備するために 今後 3年間 15億ドルの 資金が 必要なことと 報告 各国に 支援を 要請しています.
T3：日本政府は, 今回の 会合で 支援を 実施する 前提にで, 軍事費 削減と 税金 徴収体制の 強化で財政状況を 改善することと ODA/政府開発院jawを めぐった 取りあえずを 根絶するように 要求する 方針です.
T4：また, 保健医療の 改善と 道路建設のため 無償援助など 今年度に 約 1億 ドルの 援助を 実施して, カンボジアの 復興を 支援することを 表明する 方針です.

図３　図２の韓国語記事を機械翻訳システムで翻訳した結果

スを聞く人にとってわかりやすい表現が用いられる. 例えば, 経済ニュースの日本語記事を英語記事に翻訳するときには「円からドル表記へ」の変換が行われる. したがって, 我々の対象とする記事では単純に数字の一致による対応付けができない場合もあり, 他の情報が必要となる.

そこで, 本手法では単語の位置という情報に注目し, これを利用して記事の自動対応付けを試みた. 翻訳の際には名詞連続のように, 単語が出現する順序が保存される場合も多い. また, 「動詞とその目的語となる名詞」のように, 言語によってはその順序がかわるものの, 近接して出現する場合も多い. 図３は図２の韓国語記事を市販の機械翻訳システム（(株) 高電社のJ-Seoul）を使って翻訳した例であるが, 日本語と韓国語では言語的に非常に近いために, 翻訳された記事の単語の順序は元の日本語記事と非常に似ている. このような単語の出現順序を記事の自動対応付けに利用することにより, 対応付けの精度を向上させることが期待できる.

3　単語出現位置による自動対応付け

3．1単語出現位置

単語出現位置とはある単語の記事中での出現場所を表したものであり，単語出現位置辞書とは記事中のすべての単語に対して単語出現位置を集めたものである[加藤 99]. ここで，出現場所は記事番号（4桁），文番号（2桁），単語番号（2桁）を順に並べることによって定義している．例えば，“カンボジア”の単語出現位置が「06370111」とは，単語“カンボジア”が記事番号「0637」の日本語記事の，第「01」番目の文の，第「11」番目の単語であることを意味する．

提案する手法では，単語出現位置辞書は自動対応付けの候補となる記事を形態素解析して自動的に作成される．このような多言語記事は，元の記事が書かれた，当日あるいは前日の記事を参照する場合が多い．逆にこのような時間的情報を使って記事の自動対応付けする際に，対象となる元記事の範囲を絞ることができる．従来の自動対応付けでもこのような時間的制約が利用されておりその有効性が確認されている[高橋 97]. 提案する手法では時間制約は単語出現位置辞書を作成する際の記事候補の選択として反映されている．後述する評価実験では，翻訳される記事は当日の日本語記事か前日の日本語記事が多いという事実に基づき，今回は前々日までの3日間の日本語記事（約750記事）の中から翻訳元の記事を探した．

3．2提案手法

以下では韓国語記事と日本語記事の自動対応付けを例に取り説明する．

本手法では，機械翻訳を使って韓国語の記事を日本語に翻訳し，複数ある日本語記事の候補の中から翻訳結果（日本語）に一番類似した記事を選択する．日本語記事候補を選択する際には単語出現位置を使って翻訳した単語が出現する箇所を探す．

本手法の詳細を図4に示す．

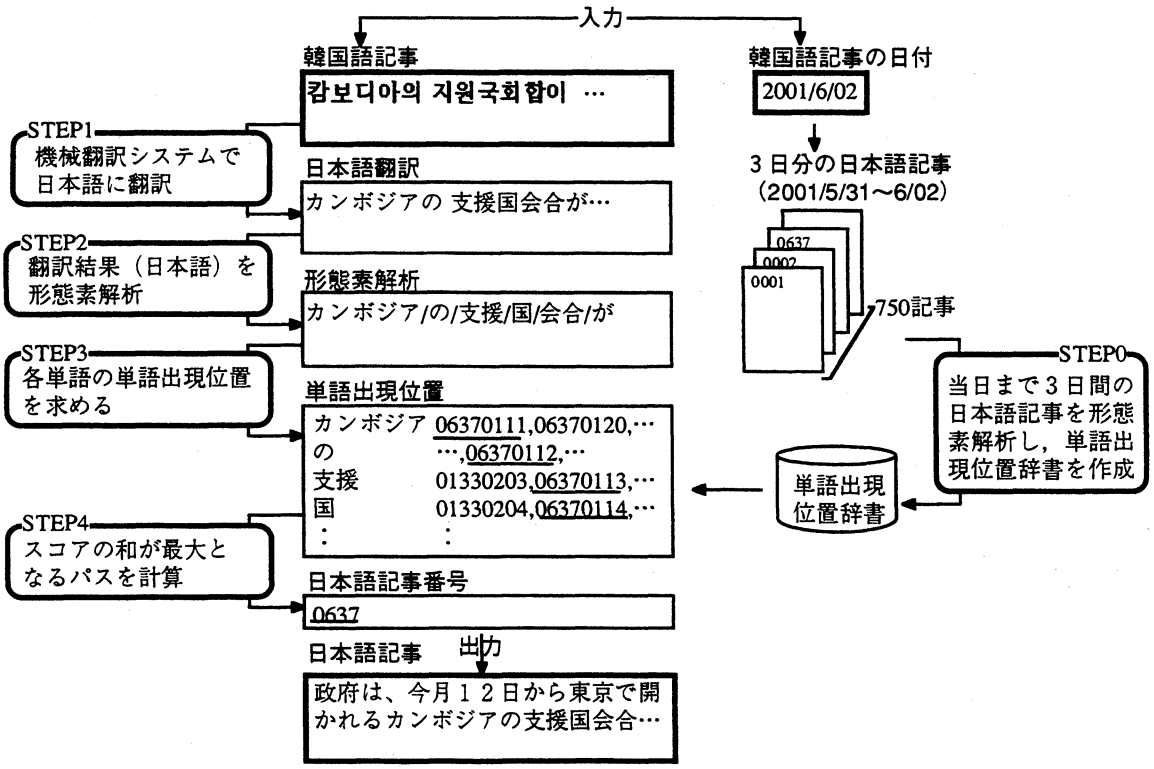

図4　機械翻訳を用いた韓国語・日本語記事の対応付け

STEP0では，その韓国語記事が書かれた日までの3日間の記事から，単語出現位置辞書を作成する．

STEP1で，機械翻訳を使って韓国語の記事を日本語に翻訳する．韓国語と日本語のように非常に近い言語では語順も保存される．

STEP2では得られた翻訳結果を形態素解析する．今回はさまざまなシステム（例えば市販システム）を使うことを前提にしてSTEP1とSTEP2を分けているが，もちろん機械翻訳システムの結果として単語の分割点が出力されるのであれば形態素解析の必要はない．

STEP3では単語出現位置を使って各単語の日本語記事中の出現位置を求める．

STEP4では，単語出現位置が，STEP2で得られた日本語の各文に対して連続する単語列$w_i w_{i+1} \cdots w_{i+n-1}$のスコア（**単語列スコア**）を計算し，1文中で単語列スコア$score(w_i w_{i+1} \cdots w_{i+n-1})$の総和を求める．ここで，単語列スコアは，単語出現位置が連続している長さをnとすると

$$score(w_i w_{i+1} \cdots w_{i+n-1}) = 3^n \quad (1)$$

と定義した．例えば，図4の例では単語列「カンボジアの支援国」の単語列スコアは，単語出現位置が06370111～06370114と4つ連続しているので，記事番号06370に対して3^4となる．最後に，記事中の全ての文に対して，スコアの和を求め．最大となる記事番号を対応する日本語記事として出力する．

4　評価実験

提案手法を使って韓国語・日本語記事，中国語・日本語記事の自動対応付けをする実験を行った．自動対応付けの対象となる韓国語記事は6,611記事，中国語記事は6,975記事であり，韓日機械翻訳，中日機械翻訳にはそれぞれ（株）高電社の「j·Seoul」，「j·北京」を使った．結果を表1に示す．ただし，precisionは対応付けがされた結果から100記事をランダムサンプルして計算した値である．

表1　自動対応付け結果

言語ペア	precision	recall
韓国語・日本語	100	94.3
中国語・日本語	76	68.6

表1を見ると，韓国語·日本語記事対応付けはprecision，recallともに非常によいことがわかる．これは日本語記事から韓国語記事を作成する際にほぼ忠実に訳されており，単語の出現順を利用したことが有効に働いているためであると考えられる．一方，中国語·日本語記事対応付けは精度があまりよくない．これは中日機械翻訳の精度が韓日ほどよくないことも原因ではあるが，韓国語ほどには単語の出現順が保存されていないことも原因であると考えられる．

5　おわりに

単語が出現する位置（単語出現位置）を利用して二言語間の記事対応をつける手法について述べた．今回は記事対応のみについて述べたが，文位置も利用しているので，実際には記事対応と同時に文対応も付けられている．今後は文対応付けの評価についても考えていきたい．

本手法は機械翻訳の性能，特に訳語選択の性能によって精度が左右される．これは機械翻訳では訳語が1つに限定されてしまうためである．そこで，対訳辞書を使うことで精度が向上するものと思われる．

【参考文献】

[Hasan 01] Md Maruf Hasan and Yuji Matsumoto "Multilingual Document Alignment - A Study with Chinese and Japanese.", Proc of NLPRS2001, pp.617-623, 2001.

[加藤 99]加藤ほか「ニュース音声認識のための(n≧4)-gramを併用する言語モデル」電子情報通信学会研究報告，SP99-125, pp.55-60, 1999.

[松本 01]松本賢司，柏岡秀紀，田中英輝「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」情報処理学会第63回全国大会，Vol.2, pp.251-252, 2001.

[高橋 97]高橋大和，白井諭，大山芳史「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，pp.127-130, 1997.