類似した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
{katou.n-ga, ehara.t-eo}@nhk.or.jp

1　はじめに

機械翻訳や翻訳メモリによる翻訳支援システムを構築する場合，大規模な対訳コーパスが重要となってきている．対訳コーパスを収集する一つの手段として，文書の対応情報が失われてしまった２つの単言語コーパスから，自動的にその対応を求める手法が提案されている[1,2,3,4]．その多くは対訳辞書を利用している．特に，日本語と英語のように言語的に遠い言語ペアの場合には対訳辞書は不可欠である．しかし，対訳辞書は言語ペアによっては入手困難であるという問題がある．一方，ポルトガル語とスペイン語のように同族の言語ペアの場合には，対訳辞書を使わずに言語的特徴のみに基づき対応付けすることが考えられる．実際，文書の対応付けではないが，文や単語などの対応付けに，言語の表層的な特徴が利用されている[5]．

本稿では，このような同族語がもつ言語類似性を使って，多言語のニュース記事を効率的に自動対応付けする手法について述べる．文書の自動対応付けは，文や単語の対応付けに比べ，処理をする単語の数が非常に多くなるが，本手法では，文字出現位置辞書，単語出現位置辞書を使うことにより，効率よく類似箇所を発見している．

2　多言語放送ニュース記事

図１に我々の対訳コーパスの元となる，多言語放送ニュース記事が作成される流れを示す．まず，日本語ニュース記事から英語ニュース記事が独立に作成される．次にこの英語ニュース記事を元に，ポルトガル語，スペイン語，‥とそれぞれのニュース記事が作成される．したがって，英語とポルトガル語のニュース記事には直接の対訳関係があるが，例えば，ポルトガル語とスペイン語のニュース記事には直接的な対訳関係はない．しかし，ポルトガル語においてもスペイン語においても，元の英語ニュース記事を忠実に翻訳している部分も多く，後述する例を見てもわかるように，対訳として利用可能なものも多い．図２はそのようなポルトガル語，スペイン語，イタリア語，ドイツ語，スウェーデン語，英語の６言語のニュース記事の例である．各ニュース記事は，それが書かれた**言語**(LANG)，**日付**(DATE)，**タイトル**(TITLE)，**文書番号**(FILE)，**本文**(S)から構成されている．タイトルは，翻訳元である英語のニュース記事に付与されているものを他の言語にもコピーしている．このタイトルを利用すれば，各言語間のニュース記事の対応を付けることができる．しかし，タイトルは現在人手で入力しているために入力もれもあり，実際には対応関係をつけることができない場合も少なくない．次に本文を見ると，対訳とみなせる箇所があることがわかる．例えば，図２において，ポルトガル語の第１文 “A ministra da Educação e Ciencia do Japáo ...” と，スペイン語の第１文 “La Ministra de Educación y Ciencias de Japón ...” はおおむね対訳とみなせる．そしてこの２文には表層上の類似性もある．例えば，単語上では，“A” と “La”，“ministra” と “Ministra”，“da” と “de”，“Educação” と “Educación”，... というようにアルファベットが類似している．また，その単語の並びすなわち構文的にも類似している．これはポルトガル語とスペイン語が同族であることによる．本稿で述べるニュース記事の自動対応付け手法は，このような同族語がもつ，単語類似性と構文類似性を使っている．

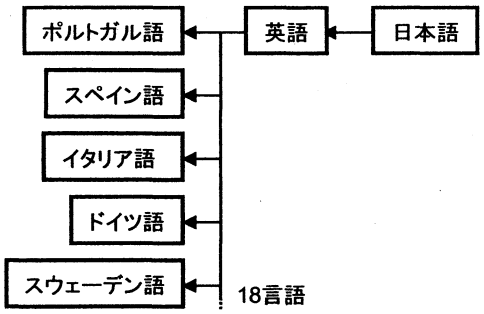

図１　多言語ニュース記事作成の流れ

LANG:ポルトガル語
DATE|2001/07/24
TITLE|06 Toyama to attend Perus presidential inauguration
FILE|Po0016
S|A ministra da Educação e Ciencia do Japáo, vai representar o govenro japonès na cerimonia de posse do novo presidente do Peru, nesta semana.
S|A cerimonia de posse do presidente Alejandro Toledo vai acontecer no sábado. Ele venceu a eleiáo presidencial em seguida à dmeissáo do presidente Alberto Fujimori.
S|A Ministra Toyama deverá partir do Japáo na quinta-feira e vai manter conversações com Toledo antes mesmo da cerimonia de posse.
S|Noso reporter indica que o govenro decidiu enviar Toyama ao Peru numa tentativa de confirmar os laços de amizade existentes entre os dois píases a depseiot de questões pendentes que estáo pressionando esta situação.
S|A promotoria pública do Pru está solicitando que o Japáo faça a extradição do ex embaixador perunao no Japáo, Victor Aritomi, que é cunhado de Fujimori.
S|Uma ordem de prisáo foi emitida contra ele sob a acusaão de envolvimento no ecandalo de desvios de fundos pelo ex-presidente peruano.
S|Entretanto o Japáo decidiu dar a cidadania japonesa para Aritomi , recusando assim a solicitação de extardição feita pelo Peru.

LANG|スペイン語
DATE|2001/07/24
TITLE|06-06 Toyama to attend Peru's presidential inauguration
FILE|Sp0008
S|La Ministra de Educación y Ciencias de Japón, Atsuko Toyama, representará al gobierno japonés en la toma posesión de Alejandro Toledo como nuevo presidente de Perú este sábado.
S|Alejandro Toledo venció a su contrincante Alan García en una apretada segunda ronda electoral celebrada el tres de junio tras la destitución del ex presidente Alberto Fujimori, que abandonó repentinamente el cargo en octubre para refugiarse en Japón, donde se encuentra actualmente.
S|La ministra Toyama planea partir de Japón el jueves y sostener conversaciones con Toledo antes de su toma de poder.
S|Nuestro reportero comenta que el gobierno ha decidido enviar a Toyama a Perú en un intento por confirmar los vínculos amistosos entre ambos países a pesar de ciertos asuntos que han empañado las relaciones bilaterales.
S|La fiscalía peruana exige a Japón que extradite al ex embajador de Perú ante Japón, Victor Aritomi, cuñado de Fujimori.
S|Se ha emitido una orden de arresto en contra de Aritomi por su presunta vinculación en un escándalo de soborno del ex presidente.
S|Sin embargo, Japón decidió naturalizar a Aritomi y rechazar la exigencia peruana.

LANG|イタリア語
DATE|2001/07/24
TITLE|06-06 Toyama to attend Peru's presidential inauguration
FILE|It0001
S|Il Ministro della Pubblica Istruzione e della Scienza Atsuko Toyama rappresenterà il governo giapponese alla cerimonia per l'insediamento ufficiale del nuovo Presidente peruviano.
S|La cerimonia per il neoeletto Presidente Alejandro Toledo si terrà sabato. Toledo ha vinto le elezioni inseguito alle dimissioni dell'ex Presidente Alberto Fujimori.
S|I magistrati peruviani stanno richiedendo l'estradizione dal Giappone dell'ex ambasciatore a Tokyo Victor Aritomi.
S|Un mandato di arresto è stato rilasciato per il suocero di Fujimori accusato di essere rimasto coinvolto nello scandalo per corruzione dell'ex Presidente.
S|Tuttavia il Giappone ha deciso di dare ad Aritomi la nazionalità giapponese e ha rifiutato la richiesta da parte peruviana.

LANG|ドイツ語
DATE|2001/0724
TITLE|06-06 Toyama to attend Peru's presidential inauguration
FILE|Ge0005
S|Die japanische Ministerin für Erziehung und Wissenschaft wird an den Amtseinführungsfeierlichkeiten für den neuen peruanischen Präsidenten in dieser Woche teilnehmen.
S|Die Zeremonie für den designierten Präsidenten Alejandro Toledo findet am Sonntag statt. Toledo hatte die Präsidentschaftswahlen nach der Amtsenthebung des Ex-Präsidenten Alberto Fujimori gewonnen.
S|Frau Toyama wird Japan am Donnerstag verlassen und vor den Feierlichkeiten zu einem Gespräch mit dem designierten Präsidenten zusammenkommen.
S|Wie ein NHK-Reporter erklärte, habe Tokio die Ministerin nach Peru entsandt um die freundschaftlichen Beziehungen zwischen Japan und Peru zu unterstreichen trotz der jüngsten Spannungen in den Beziehungen.
S|Der peruanische Regierung fordert von Tokio eine Auslieferung des ehemaligen peruanischen Botschafters in Tokio, Victor Aritomi.
S|Ein Verhaftungsbefehl liegt gegen den Schwager von Ex-Präsident Alberto Fujimori mittlerweile vor.
S|Tokio hatte jedoch entschieden, den Antrag von Aritomi auf japanische Staatsbürgerschaft anzuerkennen und lehnt daher die Auslieferungsforderungen ab.

LANG|スウェーデン語
DATE|2001/07/24
TITLE|06-06 Toyama to attend Peru's presidential inauguration
FILE|Sw0002
S|Utbildningsminister Atsuko Toyama kommer att representera den japanska regeringen vid tillsättningen av Perus nye president den här veckan.
S|Invigningsceremonin för den tillträdande presidenten Alejandro Toledo kommer att ta plats på lördag.
S|Han vann presidentvalet efter avsättningen av Alberto Fujimori från posten.
S|Dom peruanska åklagarna kräver att Japan utlämnar landets före detta ambassadör till Tokyo, Victor Aritomi.
S|En häktningsorder utfärdades för Aritomi, som är svåger till Fujimori och står anklagad för involvering i den före detta presidentens avlönings-skandal (?).
S|Men Japan beslöt att ge Aritomi medborgarskap och avslå den peruanska begäran.

LANG|英語
DATE|2001/07/24 01:43
TITLE|Toyama to attend Peru's presidential inauguration
FILE|En0092
S|Education and Science Minister Atsuko Toyama will represent the Japanese government at the inauguration of the new Peruvian president this week.
S|The inauguration ceremony of President-elect Alejandro Toledo will take place on Saturday.
S|He won the presidential election following the dismissal of former President Alberto Fujimori.
S|Minister Toyama is to leave Japan on Thursday and have talks with Mr. Toledo before his inauguration.
S|NHK's reporter says the government decided to dispatch Ms. Toyama to Peru in an attempt to confirm friendly ties between the two countries despite issues straining bilateral relations.
S|The Peruvian prosecution is demanding that Japan extradite Peru's former ambassador to Tokyo, Victor Aritomi.
S|An arrest warrant was issued for the brother-in-law of Mr. Fujimori on charges of involvement in the former president's payoff scandal.
S|However, Japan decided to grant citizenship to Mr. Aritomi and refuse the Peruvian demand.

図２　多言語放送ニュース記事の例
（ポルトガル語，スペイン語，イタリア語，ドイツ語，スウェーデン語，英語）

3　言語類似性による文書対応

本手法の全体図を図３に示す．本手法では，２つの文書（**対訳元**と**対訳先**のニュース記事）の自動対応付けを行う際に，まず**単語類似度**に基づいて，対訳先候補（対訳元が書かれた当日と前日のニュース記事）に出現する類似単語を抽出する．次に，**構文類似度**により，単語の並びがもっとも似ている部分が多い文書を最終的な対応先として求める．単語類似度，構文類似度はそれぞれ，**文字出現位置辞書**，**単語出現位置辞書**を使って計算している．文字出現位置辞書は単語出現位置辞書から，単語出現位置辞書は対訳先候補の文書から自動的に作成される．以下ではそれぞれを説明する．

３．１　単語類似度

単語類似度は，McEnery ら[8]と同様に，２つの単語中一致する文字数からDice係数で計算している．ただし，本手法では，２文字以上連続して一致する文字列を優先し，そのような文字列は文字出現位置辞書を利用して求めている．文字出現位置辞書とは，ある文字（アルファベット）が後述する単語出現位置辞書中での**文字出現位置**（どの単語のどこに出現しているか）を集めたものである．文字出現位置は７桁の数字XXXXYYY(XXXXは単語出現位置辞書のエントリー番号，YYYはその文字の単語の先頭からの順番）で構成されている．例えば，対訳元の“ministra”は，２文字目以降の文字“inistra”の文字出現位置が0211002(i)，0211003(n)，...,0211008(a)と連続しているので，単語出現位置辞書のエントリー番号０２１１の単語“Ministra”と７文字一致していることがわかる．このとき，“ministra”と“Ministra”の単語類似度は(2*7)/(8+8)となる．同様にして，図３のように，対訳元の単語“ministra”に類似した，対訳先の単語が他に“ministra”，“Ministro”と求めることができる．

３．２　構文類似度

構文類似度は，単語列の並びが連続して一致する箇所のスコアを3^n（n は連続して一致する単語数）として，２つの文中の総スコアで計算している．このとき，単語列の一致は単語出現位置辞書よ

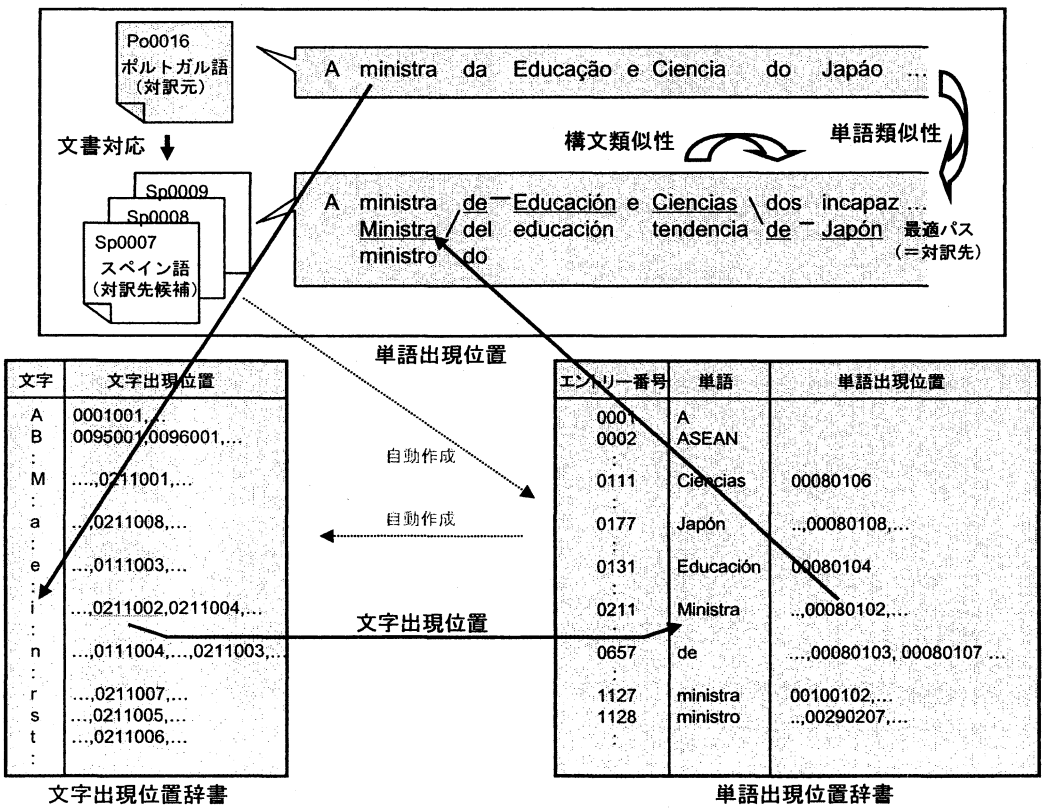

図３　文字出現位置辞書と単語出現位置辞書による文書対応

り求めている．単語出現位置辞書は対訳先候補中の**単語出現位置**（どこの文書に出現しているか）を表したものである．単語出現位置は８桁の数字ZZZZWWVV（ZZZZはファイル名，WWは文番号，VVはその単語の文頭からの順番）で表している．例えば，“Ministra”は文書番号0008の第１番目の文の２番目に出現しているので00080102となる．

4　実験

本手法を用いて，ポルトガル語のニュース記事から，スペイン語，イタリア語，ドイツ語，スウェーデン語，英語のニュース記事への自動対応を行った．各言語の文書の特徴を表1に示す．実験に用いたニュース記事は，2000年４月から2002年４月までの記事の中で，タイトル等を使って対応付けができたものである．実験結果を表２に示す．

表２を見ると，ポルトガル語のニュース記事から，スペイン語，イタリア語のニュース記事への自動対応の精度が非常によいことがわかる．次いで，英語，スウェーデン語に対する精度がよく，これらに比べてやや低くなるのがドイツ語である．これらの精度の違いは言語の近さによるところが大きい．この比較を言語の分類をしている言語系統樹[7]と比較してみると，ポルトガル語，スペイン語，イタリア語はイタリック語派としてまとめられており，実験結果でも非常に正解率が高い．一方，英語，スウェーデン語，ドイツ語はゲルマン語派に分けられ，ポルトガル語より言語的にやや遠い．ゲルマン語派はさらにドイツ語と英語の西ゲルマン語と，スウェーデン語の北ゲルマン語に分かれる．この分類からすると，ドイツ語と英語の正解率が同じ程度になることが予想されるが，今回の実験では正解率にかなりの差がある．この原因の一つには，ドイツ語では翻訳の際に意訳していることが考えられる．例えば，図２を見ると，ポルトガル語の第１文に出現している単語“representaro”が，スウェーデン語では“representera”と，英語では“represent”と現れているのに対し，ドイツ語では“teilnehmen”が使われており，単語の類似性が少ない．

表１　多言語ニュース記事の特徴

言語	一日あたりの平均記事数	一記事あたりの平均総単語数	一記事あたりの平均異なり単語数
ポルトガル語	18．8	134．0	89．0
スペイン語	19．8	130．6	83．0
イタリア語	9．8	127．8	89．0
ドイツ語	14．4	116．0	86．8
スウェーデン語	8．7	101．3	75．3
英語	30．6	148．9	96．3

表２　ポルトガル語から他の言語への自動対応付けの実験結果

言語	実験記事数 (a)	正解記事数 (b)	正解率 (b/a)
スペイン語	1，769	1，735	98．1%
イタリア語	2，035	2，011	98．8%
ドイツ語	2，589	1，901	73．4%
スウェーデン語	1，384	1，165	84．2%
英語	7，168	6，291	87．8%

5　おわりに

同族語の２つの言語を対象にして，対訳辞書を使わずに言語上の類似性に基づいてに文書対応を効率的行う手法について述べ，実験によりその有効性を確認した．

今後は単語類似度を計算する際に，大文字小文字を区別しない，先頭の１文字の一致を利用する，他の類似度[8]を利用することも考えたい．また，他の同族語言語ペア（例えば，英語ードイツ語，英語ースウェーデン語，ヒンディ語ーウルドゥ語）についても実験を行ってみたい．

参考文献

[1] 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．

[2] 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．

[3] Hasan, M.M. and Matsumoto, Y. Multilingual Document Alignment - A Study with Chinese and Japanese. In Proceedings of NLPRS2001, pp.617-623, 2001.

[4] 加藤直人，江原暉将．二言語間の放送ニュース記事の自動対応付け．言語処理学会第８回年次大会，pp．45-48，2002．

[5] Véronis, J. Parallel Text Processing. Kluwer Academic Publishers, 2000.

[6] McEnery, A.M. and Oakes, M.P. Sentence and word alignment in the CRATER project : methods and assessment. In Proceedings of the EACL-SIGDAT Workshop, 1995.

[7] 市河三喜，高津春繁[編]，“世界言語概説”，研究社．2000．

[8] 北研二，“確率的言語モデルに基づく多言語コーパスからの言語系統樹の再構築”，自然言語処理，Vol．4，No．3，pp．71-82．1997．