Webからの日英同一話題文書の抽出

濱岡 俊介 †　　馬 青 †　　村田 真樹 ‡

† 龍谷大学大学院理工学研究科数理情報学専攻

‡ 鳥取大学大学院工学研究科情報エレクトロニクス専攻

1　はじめに

われわれは単語レベルとフレーズレベルでの英作文支援システムを開発してきた．その中でわれわれはフレーズレベルでの支援の性能向上を図るため，日英対訳パターンに基づくアプローチが必要と考え，そのパターン辞書の作成に必要となる日英対訳表現を大規模な日英対訳コーパスから抽出することを試みてきた [1]．言うまでもなく，対訳表現をより大規模高精度に獲得するためには，超大規模で高精度な対訳コーパスが必要である．

そこでわれわれは，Webから日英対訳コーパスを作成することを考えた．Webから対訳コーパスを作成するには二つのフェーズが必要である．一つは対訳となりえそうな日英の文書集合を獲得すること (フェーズ1)，もう一つは獲得した文書集合を整列 (アライメント) すること (フェーズ2) である．本稿は，最初のフェーズである対訳となり得そうな日英文書集合の獲得について報告する．

対訳となり得そうな日英文書集合の獲得において，文書間に何の関連もない文書集合を獲得することは効率が悪い．そこでわれわれは，ニュース記事のタイトルから獲得された日本語の重要語と，その重要語を機械翻訳で英語に翻訳したものをそれぞれクエリとし検索エンジンから文書集合を獲得した．このように集めた文書集合に対し，話題語抽出の研究 [2] を参考にクラスタリング手法を改良し話題に分類する実験を行った．そして，文書はどのように話題に応じて分類されたか，また分類されたクラスタ内の日英文書の比率はどうであったかについて評価を行った．

2　対訳コーパスの自動作成

2.1　関連研究

対訳コーパスは，近年の言語処理研究にとって重要な資源の一つである．そのため，対訳コーパスの自動作成に対して期待は大きい．前節に述べたように，対訳コーパスをWebから自動的に作成するには二つのフェーズが必要である．フェーズ1の関連研究としてはWebから英語アラビア語間のコーパスを作るため，URL規則やHTMLのタグ情報で文書を獲得し整列させた研究 [3] やWebから対訳テキストを言語横断検索 (CLIR) によって獲得する研究 [4]，さらにはCLIRにより文書を自動獲得し訳語対応を獲得する研究 [5] などが挙げられる．一方，フェーズ2の関連研究としてはオープンソースのソフトウェアマニュアルから対訳コーパスを作成する研究 [6] やテキストが対訳となっているかを高速判定する研究 [7] などが挙げられる．

しかし，どの研究を見てもWebから自動的に反復して大量の対訳コーパスを作成するような研究は存在しない．例えば，オープンソースのソフトウェアマニュアルから対訳コーパスを作成する研究では，文書獲得はあらかじめオープンソースのマニュアルと限定しており，研究の中心は獲得文書の整列である．またテキストが対訳となっているかを高速判定する研究では，Webからどのように文書を獲得するのか明記されておらず，現実のWebデータから具体的に文書を獲得する方法について明記されていない．英語アラビア語間の研究やCLIRで文書を獲得する研究では両方のフェーズを一貫して行っているが，前者は対象が英語アラビア語間であり，後者はHTMLタグ情報やURL規則などを利用しているため大規模な対訳文書の収集に限界があるように見える．このようにWebから未知の文書集合を大量に取り出し，日英対訳コーパス自動作成へと繋げていく研究や手法がないのが現状である．

2.2　対訳コーパス自動作成の手法

われわれの最終目標はWebから日英対訳コーパスを自動作成することである．そのためには，日英の文書集合を獲得し整列する必要がある．われわれの理想とする手法は図1のようなものである．あるクエリを

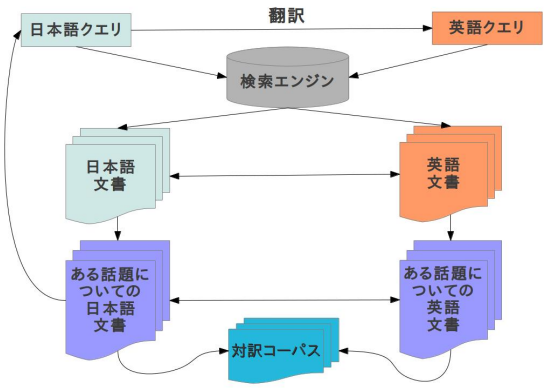

図 1: 大規模対訳コーパス作成手法

シード，すなわち，種として与えれば文書集合を自動収集する．それらの文書を話題によって分類し，文レベルでアライメントする．次なるクエリは収集された文書から自動作成し，それを用いて上記同様な処理を行う．このような繰り返し処理で対訳コーパスを自動作成していく．

今回は上記処理中の，シードを与えて文書集合を収集し，話題に分類する研究を行った．

3　日英文書の獲得

図 1 の最初のステップである日本語クエリと英語クエリの獲得，及び日本語文書，英語文書の獲得について述べる．一番最初のクエリは複数の話題の元となり得そうな文書がよいと考えた．

まず，日英共に話題となりそうなニュース記事をニュースサイトから選び，その記事からタイトルを抽出する．そのタイトルに対して Yahoo!Japan の提供する Yahoo!キーフレーズ抽出 [8] を利用し，キーフレーズを獲得する．これを日本語クエリとし，検索エンジンの AND 検索で日本語の文書集合を獲得する．Yahoo!キーフレーズ抽出を利用することにより，他の手法では困難な固有表現の処理が簡単にできる．一方，英語クエリは，日本語クエリから Google 翻訳 [9] を利用して得られたものである．クエリの翻訳は辞書引きでも良いが，未知語への対応などから Google 翻訳を使用した．このようにして得られた英語のクエリを用いて検索エンジンの AND 検索で英語の文書集合を獲得する．

4　日英文書混在のクラスタリング

4.1　考え方

日英のクエリで関連性の高い文書集合を獲得した後は，文書同士を効率的に整列するためにさらに文書同士を話題の関連度で絞り込みたい．そこで話題語獲得の手法 [2] を参考に，改良したクラスタリング手法を用いて，文書集合の話題による分類を試みる．ただし，英語の文書を日本語の文書と混在した状態でクラスタリングを行うため，Google 翻訳を利用して英語の文書を日本語のものに変換する．これにより日英混在の文書集合に対してクラスタリングを行うことが出来る．このようなクラスタリングを行うことにより，同じクラスタ内に日本語と英語の文書が所属すればその文書同士は同一の話題について述べられている可能性が高いと期待できる．

4.2　クラスタリング

文書 i のベクトルは以下のように構成する．

$$d_i = (x_{i1}, x_{i2}, \cdots, x_{iV}) \tag{1}$$

ただし，V は文書集合全体における名詞の数であり，数，副詞可能，代名詞などは取り除いてある．また，x_{iv} は以下のように求める．

$$x_{iv} = \begin{cases} \log\{M/df(w_v)\} & \text{if } tf(d_i, w_v) \neq 0 \\ 0 & \text{else} \end{cases} \tag{2}$$

ただし，$df(w_v)$ は全文書中で単語 w_v が出現する文書の数，$tf(d_i, w_v)$ は文書 d_i における単語 w_v の出現回数，M は以下のように求める．

$$M = \max_{v=1,\cdots,V} df(w_v) \tag{3}$$

文書 i, j 間の類似度 S_{ij} は以下のように求める．

$$S_{ij} = \frac{d_i \cdot d_j}{|d_i||d_j|} \tag{4}$$

クラスタリングは文書間の非類似度 (1　S_{ij}) を文書間距離として行う． ここでは先行研究と同様，最長距離法を利用するが，処理の終了を距離でなくクラスタ内に属する文書数で判定する．クラスタ内に所属する文書数を用いることにより，話題分割の細かさを直感的にコントロールすることが出来る．また，最初のクラスタを文書獲得に使ったニュース記事の本文（以降「シード記事」と呼ぶ）にすることで初期値によるクラスタリングの結果のブレを無くす．

クラスタリングが終了した後のクラスタ評価に，三つの尺度を用いる．クラスタ内の文書数，シード記事

とクラスタ中心との文書類似度，そしてクラスタの密度である．クラスタ中心との文書類似度を求めるためには先行研究に用いた文書話題度を用いてクラスタ内の中心文書を割り出す必要がある．一方，クラスタの密度は，クラスタ内の文書間の類似度の総和を文書数で正規化したもので定義する．以下，それらの求め方を示す．

まず，文書話題度を以下のように求める．

$$DocTopic(i) = \frac{1}{N_C} \sum_{j \in C,\ i \neq j} S_{ij} \qquad (5)$$

ただし，i は文書番号，C はクラスタ番号，N_C はクラスタ C 内の文書数である．そして，文書話題度を用いてクラスタ C の中心文書は次のように求める．

$$CentralDoc(C) = \arg\max_{i \in C} DocTopic(i) \qquad (6)$$

また，クラスタの密度は以下のように求める．

$$ClusterDensity(C) = \sum_{i \in C} DocTopic(i) \qquad (7)$$

5 実験

5.1 概要

朝日新聞のニュースサイトからタイトルが「ＮＺクライストチャーチ市、地震に強い街づくり案を発表」のニュース記事をシードとして選び，タイトルから日英の文書集合の取得を行った．表 1 はタイトルから獲

表 1: 日英クエリ

日本語のクエリ	英語のクエリ
NZ クライストチャーチ市 強い街づくり案 発表 地震	City of Christchurch NZ Intense urban development plan Announcement Earthquake

得された日英のクエリであった．獲得された文書は，日本語の文書上位 100 件，英語の文書上位 100 件の計 200 件である．獲得件数については，図 1 に示した手法では何件でもよいが，今回は数多くの話題を獲得するため，データを多めに取得した．また，シード記事として地震が起きてから派生した話題のものを選んだ．これは，偏った話題からでも話題を獲得できるか確かめるためであった．

獲得された文書集合に対して改良したクラスタリングを行った．初期クラスタをシード記事の本文にした．また，クラスタリングの終了条件を，クラスタに属する文書数が 30，20，10 以下という三つの場合とした．文書ベクトルの単語数 V は 26,900 であった．

5.2 結果及び評価

5.2.1 話題の定義

話題は，日英文書 200 個の中に含まれる話題一つ一つのことを言う．今回は実験の中での話題を「クライストチャーチ市での地震」に関連する話題と定義する．例えば「パーカー市長が緊急声明を発表」と言うのが例である．

今回獲得された日英合わせて 200 文書の中には大きく分けて話題は 30 個あった．表 2 は代表的な話題を示す．

表 2: 代表的な話題

クライストチャーチ市で地震発生
パーカー市長が緊急声明を発表
復興総額は 120 億ドル以上かかる模様
本震のマグニチュードは 7.1
留学生含めて 180 名が犠牲に
募金総額が 1030 万 NZ ドルに
シンボルである聖堂が崩壊

5.2.2 選択されたクラスタ

表 3 は三つの異なる終了条件でのクラスタリング結果に対して，4.2 節に述べたそれぞれのクラスタ評価尺度で選択された Top 3 のクラスタ内の話題数を示す．/で区切られた数字は左から順に 1 位，2 位，3 位を示す．

表 3: 各 Top 3 のクラスタ内の話題数

終了条件	文書数	密度	クラスタ中心との類似度
30	8/4/5	0/0/3	2/5/5
20	3/4/3	4/0/1	2/5/12
10	13/1/1	0/3/0	5/0/2

表 3 から，終了条件であるクラスタ内の文書数が少なくなるに連れて話題の数が減ることがわかる．しかし，終了条件が文書数 20 で中心文書との文書類似度が 3 位のクラスタや，終了条件が文書数 10 で文書数 1 位のクラスタが極端に多い話題を持ってしまっている．これは上位クラスタが変わったことや新たに所属した文書内に話題が多く含まれていたことなどに起因すると見られる．

5.2.3 除外されたクラスタ

クラスタ内の文書数が少なく，クラスタ中心との文書類似度の低いクラスタは除く．具体的な閾値として今回は文書数を 3 以下でクラスタ中心との文書類似度

が 0.0003 以下のクラスタ及びクラスタ内の文書を取り除くことにした．結果，それらのクラスタやクラスタ内の文書を合わせると 23 個の文書が除外された．これら 23 個の文書の中にはトータルで話題が一つしか含まれていなかった．この結果，データ数が少なく，クラスタの中心文書がシード記事とあまり類似しないクラスタを取り除くことは妥当であることが確認された．

5.2.4 クラスタ内の日英の文書の数

次に各クラスタ内の日英文書の割合を示す．結果は表 4 の通りである．文書が日英のどちらかに偏っていることが多かった．これは文書を獲得する際の HTML のタグ消去や翻訳の不十分さなどにより英語の文書は英語同士，日本語の文書は日本語同士で固まってしまったと考えられる．

表 4: 文書数が最大のクラスタ内の日英文書数

終了条件	文書数	密度	クラスタ中心との類似度
30	日:3 英:26	日:0 英:2	日:1 英:0
20	日:0 英:20	日:0 英:4	日:1 英:0
10	日:1 英:7	日:0 英:3	日:1 英:0

5.2.5 IR/CLIR との比較

最後に IR と CLIR による上位検索結果との比較を行う．それぞれのクラスタリングの終了条件において，クラスタ内の文書数が最大となるクラスタ内における日本語の文書の数だけ IR 上位の結果を獲得する．同様に英語の文書の数だけ CLIR の結果を獲得する．結果は表 5 のようになった．

表 5: クラスタリングと IR/CLIR の結果比較

終了条件	クラスタ内の話題数	IR/CLIR の話題数
30	8	14
20	3	12
10	13	7

終了条件が 10 以下の場合以外はクラスタリングにより話題が絞れたのがわかる．終了条件が 10 以下の場合に話題が多いのは，一つの文書に 8 個話題が載ったものがあったためである．

5.2.6 考察及び今後の課題

クラスタ内の文書数を少なくしていくことで所属する文書数を減らし，クラスタ内で話題を絞って獲得することが出来た．しかし，一つの文書に様々な話題が載ることも多く，クラスタリングの終了条件が 10 の時の文書数最大のクラスタは多くの話題を持つ結果となった．また，日英文書の割合が期待したようにばらけなかった．これは翻訳の精度であったり，文書データの HTML タグなどの取り除き方などを改良する必要があると考えられる．

今後の課題としては，もっと多くのデータに対してクラスタリングを行い結果を確かめることなどが挙げられる．また，もっと話題を限定的に絞り込むため，話題に応じて文書ベクトル内の単語の重みを変えるなど既存の手法の改良や新しい手法の提案などが挙げられる．

6 おわりに

大規模な日英対訳コーパスを自動作成することを目標とし，その第一ステップとなる，対訳となり得そうな文書集合の獲得を試みた．日英クエリで検索して得られた日英文書を話題関連度に応じて分類するために用いた最長距離法クラスタリングにおいては，日英文書を混在にし，シード記事を最初のクラスタにおき，クラスタ内の文書数を終了条件とするように，改良した．また，クラスタの評価（選択と除外）にクラスタ内の文書数，クラスタの密度，そして中心文書とシード記事の類似度という三つの評価尺度を導入した．実験結果から，獲得した文書集合への話題絞込みにある程度有効であることがわかった．

参考文献

[1] 坂上，馬，村田 : 辞書情報と規則を用いた大規模な日英対訳表現の抽出，言語処理学会第 17 回年次大会，pp. 983-986 (2011)

[2] 佐藤，川島，佐々木，大久保 : 文書の類似度と新鮮度に基づく話題語抽出，情報処理学会自然言語処理研究会，Vol. 2005, No. 1, pp. 29-35 (2005)

[3] Resnik, Smith: The Web as a Parallel Corpus, Computational Linguistics - Special Issue on Web as Corpus, Vol. 29, Issue 3, pp. 349-380 (2003)

[4] Nie, Simard, Isabelle, Durand:Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web, SIGIR '99, pp. 74-81 (1999)

[5] 堀内，千葉，浜本 : 言語横断探索より自動収集された日英関連報道記事からの訳語対応の獲得，電子情報通信学会言語理解とコミュニケーション，Vol. 102, No. 200, pp. 93-100 (2002)

[6] 石坂，内山，隅田，山本 : 大規模オープンソース日英対訳コーパスの構築，情報処理学会自然言語処理研究会，Vol. 2009-NL-191, No.17, pp. 1-6 (2009)

[7] 福島，田浦，近山 : 対訳辞書のグラフ表現を用いた日英対訳テキストの発見，情報処理学会自然言語処理研究会，Vol. 2006, No. 36, pp. 41-46 (2006)

[8] Yahoo!キーフレーズ抽出 http://developer.yahoo.co.jp/

[9] Google 翻訳 http://translate.google.co.jp/