Web 上の既訳文書を対象とした段落アラインメント

淺利 俊介†，竹内 孔一†，阿辺川 武††，影浦 峡††
† 岡山大学大学院自然科学研究科，†† 東京大学大学院教育学研究科
†{syun1113, koichi}@cl.cs.okayama-u.ac.jp，††{abekawa, kyo}@p.u-tokyo.ac.jp

1　はじめに

現在オンラインのニュース記事などの文書を翻訳し，ホームページで公開するボランティア翻訳者の活動が活発になってきている．翻訳者が分野依存性の高い語を翻訳しようとする際，未知の専門用語，固有名詞や定型句を訳すためには既にその分野で訳されている文書を参考にする必要がある．こうした専門性の高い語は辞書を構築することは容易ではなく，また辞書としてではなくボランティア翻訳者を支援するシステムとして提供されなければ利用することが難しい．こうした背景から本研究では既訳文書対から専門用語や固有名詞，定型句の翻訳対を取り出し，ボランティア翻訳者に対して文脈まで考慮して訳語候補を提示する支援システムの構築を目標とする．

本稿は第一段階として Web 上に存在する対訳文書を対象に上記で説明した用語や句表現を抽出するための文書間のアラインメントを行う．翻訳文書のアラインメントの先行研究には内山ら [2] や宇津呂ら [3] が新聞記事の翻訳対を利用したものがあるが，これらは全て文アラインメントによる対応付けを行っている．しかしながら本研究では段落単位でのアラインメントを行う．その理由として第一にあげられるのが翻訳者が翻訳を行う時の単位は文ではなく段落であるため対応がよいことが推測されるからである．実際我々が手にしている既訳文書では段落単位では順番が逆転することは無く，また，1 対多（多対 1）のような対応は比較的少なかった．

既訳文書からのアラインメントの先行研究として品川ら [1] があるが，実験が限定的でありかつ各コストの適用が整理されていない．本研究では品川らの研究を拡張して，引用，数値，アルファベット，辞書による訳語といった多数の要素をコスト値として整理して段落間の関連度を計算する枠組みを提案する．各コスト値は基本的な絶対量が異なることから，実際の Web 上の対訳文書から対応段落を抽出する実験を行うことで，これらの異質なコスト値をどのようにまとめるかについて考察を行う．

2　システムの流れ

本システムの大まかな流れを図 1 に示す．システムに与える入力は英・日 URL 対である．まずシステムは URL 対に対応する HTML テキストをダウンロードし，Parser を用いて文書領域を抽出する．しかし文書領域には広告や訳者のコメントのみで構成された，対訳関係にない不要段落が多く存在する．そこで不要段落の多くが文章領域の先頭および後尾に塊となって存在しやすいという性質を利用し，不要段落の削除を行う．これは段落の文字数を利用することで対訳候補領域の開始と終了を決定するが，具体的な方法は品川ら [1] に譲る．そして残った対訳候補領域に対して 1 対 1 または 1 対多（多対 1）の段落アラインメントを行いシステムの出力とする．

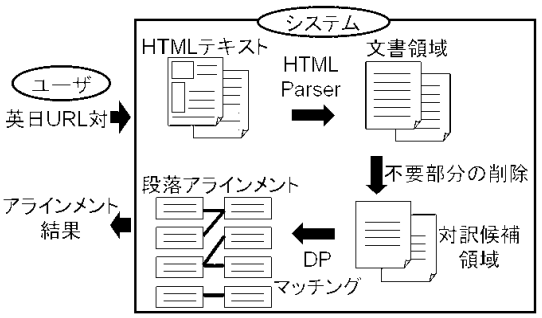

図 1: システムの流れ

3　Web 上の既訳文書の前処理

本研究では英日の既訳文書対を対象にする．Web 上にある既訳文書対から対応段落をとるには新聞記事の対訳と異なり，翻訳文と関係ない部分が存在する．これらの扱いについてここで整理することで前処理の方針を明らかにする．

3.1　特徴

翻訳者のコメント，訳注

日本語ページには翻訳者のコメントや訳注が存在する．これらのみによって構成された段落は不要段落（対訳のない段落）であるため段落アラインメントの際には対応を行ってはならない．しかしそのような段落は本文中に出現した語を使用する場合が多く，不要段落であるという判定が難しい．

広告

Web 上の既訳文書には新聞コーパス等と異なり広告が存在する．また広告の種類によっては本文の内容と関係の深いものを提示したものも存在する．広告のみによって構成される段落も翻訳者のコメントや訳注と同様に不要段落である．

文書体裁

本研究では様々なボランティア翻訳者によって書かれた Web ページを段落アラインメントの対象とする．そのためボランティア翻訳者によってカタカナや数字

の記述方法（半角，全角）など文書体裁が大きく異なることに注意する必要がある．

3.2 処理方針

翻訳者のコメント，訳注

「訳注 1：」のような翻訳者が用いる目印を元に訳注かどうかの判断を行う方法が考えられる．しかし使用する目印のバリエーションが翻訳者によって様々であるため，今回は特別な処理は行わず，段落アラインメントの際に 4.1 節で示す対応コストと，閾値となる削除コストのみによって不要段落かどうかの判定を行った．

広告

Web ページに存在する広告は，多くの場合リンクタグで囲まれている．今回はこの性質に着目し，不要段落とみなすための指標とする．具体的には「全てがリンクタグで囲まれている段落」および「リンクタグの外の部分が記号，数値表現のみで構成されている段落」については不要段落とみなし，段落対応を行わないこととした．

文書体裁

段落アラインメントで DP マッチングを行う際に，日本語ページのアルファベットを利用するため，段落アラインメントを行う前に全角アルファベットは半角に正規化を行った．同様に辞書マッチングを行う際に必要となるため半角カタカナについても全角に正規化している．また段落アラインメントでは数値表現に着目したコストも定義している．しかし日本語ページでは漢数字で数値表現を記述している場合（例：二千）がある．そこで英語ページの数値表現から日本語ページで出現する可能性のある様々な数値表現のパターンを生成し，日本語段落の数値表現とのマッチングを行った[1]．

4 段落アラインメント

文書領域候補として残った英・日の段落に対し，DP マッチングを用いて段落アラインメントを行う．この時，英・日の段落対応関係を 1 対 1 および 1 対多（多対 1）対応を考える．DP マッチングに用いるコストとして対応コストと，不要な段落の削除を行う閾値の役割を果たす削除コストを用いる．削除コストの基本的な考え方は不要段落と対訳のある段落はそれぞれある程度固まって存在する特徴を利用して求める．詳細は品川ら [1] に譲るとして，以下では各対応コストを説明した後，辞書を用いた対応する後のマッチング法について記述する．

4.1 対応コスト

英・日の段落関係を示すコストとして，文字数比，数値表現，アルファベット表記，コメント，辞書マッチ，DP マッチングで 1 対多（多対 1）対応を行う際のビーム幅に関係するそれぞれのコストを規定する．このうち文字数比を表す Character コストの最大値を 1.0 とし，数値表現，アルファベット表記，コメントに関するコストを経験的な観点から 0.0〜0.2 の範囲とする．また辞書コストについては，対訳関係にあるほどコストの合計値から引いていく．

Character コスト

日・英間での段落の文字数の比は一定であると仮定し，文字数比が α に近いものほどコストを低く定める．

$$CharacterCost = \frac{C_e - \alpha C_j}{C_e}$$

C_e は英語段落の文字数を，C_j は日本語段落の文字数を示す．

Number コスト

英語，日本語それぞれの段落中に同じ数値表現が存在する場合，両段落が対応関係にある可能性が高くなるためコストを低く定める．この数値表現の比較は 3.2 節で述べた数値表現変換により実現する．

$$NumberCost = 0.2\frac{Num_{je}}{\max(Num_j, Num_e)}$$

Num_{je} は英・日で対応している数値表現の数，Num_j，Num_e は日・英それぞれの数値表現の数を表す．

English コスト

日本語段落中にアルファベット表記が存在し，英語段落に同じ英単語が存在する場合，両段落が対応関係にある可能性が高くなるためコストを低く定める．この English コスト値の式は以下となる．

$$EnglishCost = 0.2(1 - \frac{Eng_{je}}{Eng_j})$$

ここで Eng_{je} は英・日両段落に存在する英単語の数，Eng_j は日本語段落に含まれる英単語の数である．

Comment コスト

日本語段落に「,」英段落に”，” や’，’ のようなコメント部分がそれぞれ存在する場合，両段落が対応関係にある可能性が高くなるためコストを低く定める．

$$CommentCost = \begin{cases} 0.0 & (if\ Com_{je}) \\ 0.2 & (otherwise) \end{cases}$$

Com_{je} は英・日両段落にコメントが存在する場合を示す．

Dictionary コスト

ストップワードを除く英語段落の自立語で辞書引きを行い，対訳候補が日本語段落の語と一致する場合は対応コストを下げる．

$$DictionaryCost = -\beta\frac{Dic_e}{Dic_{je}}(TFIDF_j + TFIDF_e)$$

β は Dictionary コストとその他コストの重みを決定するもので，実験では β の値を変更して段落アラインメ

[1]例えば英語ページにおける「2000」や「two thousand」について生成されるパターンは以下となる：「2000」「2,000」「２０００」「2，０００」「二〇〇〇」「二千」

ントを行っている．また Dic_e は実際辞書引きを行った英段落の自立語数，Dic_{je} は，辞書引きの結果日本語段落の語と一致した数である．$TFIDF_j$，$TFIDF_e$ は対訳関係にある日・英の語の TFIDF 値を表す．辞書には英辞郎（英語見出し：約 20 万エントリ）を使用した．

Beam コスト

1 対多（多対 1）対応の段落アラインメントを行う際に DP マッチングのビーム幅を拡張するが，このときに 1 対多（多対 1）対応は起こりにくいという観点から，Beam コストを設定する．

$$BeamCost = \gamma(BeamWidth - 1)$$

BeamWidth は DP マッチングの際のビーム幅を表す．今回は γ の値を 0.2 とし実験を行った．

4.2 英辞郎を利用した日英の対応

前節の辞書コストを計算するために既訳文書の候補段落対に対して英辞郎を利用して，単語の対応を求める．しかしながら (1) 英辞郎は人間用の辞書であるため記述に構造があること (2) 単純な品詞の対応ではうまくいかないこと (3) 書き方の異形が存在し吸収する必要があることなどの工夫が必要である．以下では，その処理について具体的に述べる．

まず英語段落中の語について TreeTagger[4] にて解析を行い，自立語の場合のみ原形で辞書引きを行う．日本語段落の語についても Sen[5] にて形態素解析を行い，英辞郎による訳語候補が日本語段落の語の原形と一致する場合を辞書マッチとする．なお日本語段落側は複数形態素[2]との一致を許している．またより多くの辞書対応が可能となるように英辞郎の訳語候補に対する補完を行っている．英辞郎の見出し例を表 1 に示す．表中の「email」の訳語のような記述形式の場合，そのまま日本語段落の語と対応付けすることは出来ない．そのためそれぞれの［,］の組み合わせを別々に対訳候補と出力できるように変更している．表 1 の「email」の例では「電子メールを送る，電子メールを送信する，電子メールを出す，E メールを送る，E メールを送信する，E メールを出す」の 6 通りを対訳候補として出力する．表中の「Iraqi」の訳語「イラクの」をそのまま日本語段落の語と対応付けしようとしても段落中の語が「Iraqi Army：イラク 軍」のような場合対応付けできない．そのため英辞郎の訳語の助詞「の」を排除し，訳語候補に「イラク」を追加している．また表中の「decennium」の訳語「10 年間」のように数値表現が含まれる場合は 3.2 節の数値表現変換を行い，日本語段落中の漢数字等にも対応している．

表 1: 英辞郎の見出し

英見出し	対訳候補
email	電子メール［E メール］を送る［送信する・出す］
Iraqi	イラクの
decennium	10 年間

5 実験および考察

4 章で提案している各コストは絶対量が異なるため，単純に積み上げて DP マッチングを行うことはできない．そこで実際の Web 上の既訳文書対に対し，Dictionary コストの配合率 β を変えて実験を行い最適なコスト値を求める．

以下，その実験内容および結果についての考察を述べる．

5.1 実験

対訳関係にある日・英の URL30 対（10 の Web サイトから 3 つずつ計 10 グループ取得）に対し段落アラインメントを行った．recall と precision はそれぞれ以下の式で求める．

$$recall = \frac{\text{正解数}}{\text{実際の正解数}} \cdot 100$$

$$precision = \frac{\text{正解数}}{\text{システムの出力数}} \cdot 100$$

ここでの正解は 1 対多（多対 1）対応の際でも全ての対応が完全一致した場合に限ったものとした．4.1 節で述べた Dictonary コストにおける β の最適値を調べるため，β =0.5，0.4，0.3，0.2，0.1 でのそれぞれの結果を表 2 で示す．

表 2: 段落アラインメント結果

β	0.5	0.4	0.3	0.2	0.1
recall	72.40	72.17	73.75	73.56	72.64
precision	57.47	57.53	58.17	57.76	56.74

また段落アラインメント結果のみの精度を判断するために，対訳候補領域の抽出の際の開始と終了を人手で行った場合の実験も行った．なお人手で行うのは対訳候補領域の開始と終了の決定だけであるため，対訳候補領域中にも広告や訳者のコメントが残っている場合がある．Dictionary コストで用いる β の値は表 2 の結果から 0.3 を使用した．実験の結果 recall は 82.21 %，precision は 82.97 %となった．

5.2 考察

対訳候補領域の抽出の際の開始と終了を人手で規定した場合の recall，precision がかなり高いことから，対訳候補領域の抽出精度を向上すべきと考えられる．また今回の実験では正解を 1 対多（多対 1）対応の際に完全一致のみと定義しているため，ほとんどが一致

[2]例えば「Vietnam」の訳語候補「ベトナム共和国」は Sen では「ベトナム/共和/国」と複数形態素で構成される．

しごく一部が一致しなかった場合でも正解とみなしていない．参考のため部分一致した場合も正解と考えた場合の recall と precision を表 3 に示す．

表 3: 段落アラインメント結果

β	0.5	0.4	0.3	0.2	0.1
recall	80.16	79.80	80.52	80.23	79.58
precision	63.39	63.32	63.45	63.00	62.11

表 3 から部分一致も正解に含むと考えた場合，recall, precision ともにかなりスコアが上昇していることが分かる．どのような部分一致が起こっているかを見てみると，ほとんどの対応関係は合っている場合が多いことが分かった[3]．よって Beam コストをうまく再定義できれば完全一致するものが増える可能性がある．

個別の段落アラインメント結果を見てみると，広告などの不要段落同士を誤ってマッチングしてしまっていたり，対応のある短い段落と不要段落とをマッチングしてしまうというミスが多い．この理由としては 1 文のみで構成されるような短い段落の対応では Comment, English，Number コストに関する語が出現しにくいことがあげられる．そこで短い 1 文の段落アラインメントを行う上では Dictionary コストが特に重要となるが，Dictionary コストの項目で示した英辞郎の訳語候補補完を行っても辞書マッチがうまく行われない場合も多い．この原因の一つに，英語段落の語の品詞が日本語側でも同じ品詞で翻訳されるとは限らないということがあげられる．これに対応するためには，英語段落の語について WordNet[6] 等を用いて品詞変換を行った後に英辞郎で辞書引きを行うなどが考えられる．同じく現在の Dictionary コストでは英・日それぞれの段落内での語の位置情報を考慮していない．そのため段落内での位置が明らかに異なるような対応関係にない語に対し，辞書による結びつけを行ってしまう場合がある．翻訳を行う際，英・日間においての文順が逆転したり複数の文がひとつの文に纏められたりする場合があるが，大きな単位である段落で見た場合，語の大まかな位置情報は大きくは変化しないと考えられる．このため辞書引きの際に語の段落での位置情報を利用することは有益であると考えられる．

また現在の Comment コストは英・日それぞれの段落のコメントの有無しか考慮していない．これにより本当は対応関係にない，段落のコメント数が大きく異なる英・日の段落を誤って対応付けしてしまった結果があった．Web 既訳文書では翻訳者が英語段落の”，”を削除して翻訳するという場合はほとんどないが，逆に強調のために「,」を使用する場合がある．そのため英語段落のコメント数と同じかそれ以上に日本語段落のコメント数が存在する場合のみ Comment コストを 0.0 とするなどの回数の考慮も有益であると考えられる．また Comment コストを Dictionary コストと組み合わせれば，英・日のコメント部分の語同士が辞書マッチした場合は両段落がより対応関係にあるといえる．これは Number，English コストについても同様に組み合わせが可能である．その場合どのような形でコスト計算を行うかを現在検討中である．

[3]例：実際の正解は一つの日段落と a,b,c,d,e,f の英段落との対応関係のとき，システムの出力はその日段落と英段落 a,b,c,d,e までの対応関係．

6 まとめ

本研究の最終目標はボランティア翻訳者に対し未知の専門用語，固有名詞，定型句に対する訳語候補の提示を文脈まで考慮して行うことである．本稿ではその前段階として Web 上の既訳文書対の段落アラインメントを行うため，英日の文字数比，数値表現，アルファベット表記，コメント，辞書マッチ，ビーム幅に関するそれぞれのコストを設定し DP マッチングにより対応を行うシステムを提案した．実際の既訳文書 30 対を対象に対訳候補領域の抽出を自動で行い段落アラインメントを行ったところ，recall は 73.75 %，precision は 58.17 %という結果であった．また対訳候補領域を人手で行ったのち段落アラインメントを行った場合，それぞれ 82.21 %，82.97 %と大きく向上している．最近では中村ら [7] の Web ページの広告判定に関する研究や韓ら [8] の Web 部分抽出に関する研究も広がっているため，今後は 5.2 節で述べたそれぞれのコストの再考慮を行うとともに，対訳候補領域を抽出する際の精度向上をはかりたい．

謝辞

本研究の一部は，日本学術振興会科学研究費補助金基盤（A）「翻訳者を支援するオンライン多言語レファレンス・ツールの構築」（課題番号 17200018）の支援により行われた．

参考文献

[1] 品川哲也，森辰則，影浦峡，“オンライン対訳文書対からのテキスト領域抽出とアラインメント”，言語処理学会第 12 回年次大会発表論文集，pp.520-523，2006.

[2] 内山将夫，井佐原均，“日英新聞の記事および文を対応づけるための高信頼性尺度”，自然言語処理 10（4），pp.201-220，2003.

[3] 宇津呂武仁，池田浩，山根正也，松本裕治，長尾眞，“対訳辞書を用いた対訳文対応および未知訳語の推定”,「自然言語処理における実動」シンポジウム論文集，pp.140-143，1993.

[4] TreeTagger，http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

[5] Sen，https://sen.dev.java.net/.

[6] WordNet，http://wordnet.princeton.edu/.

[7] 中村達也，白井清昭，“ウェブページにおける非コンテンツ領域の検出”，言語処理学会第 13 回年次大会発表論文集，pp.234-237，2007.

[8] 韓浩，徳田雄洋，“Web 部分情報抽出システムとその応用”，日本ソフトウェア科学会第 23 回大会論文集，2006.