対訳文抽出におけるレイアウト情報の利用

秋本 仁志　伊川 洋平　金山 博
日本アイ・ビー・エム株式会社　東京基礎研究所
{fallbook, yikawa, hkana}@jp.ibm.com

1　はじめに

ビジネスの国際化に伴い，サービスの提案，ソフトウェアに対する要求，技術の解説などのビジネス文書の翻訳の需要が急増している．ビジネス文書の翻訳には迅速さと均質さが求められるため，過去の翻訳結果のうち類似するものを再利用する仕組みがあるのが望ましい．一方で，日々の業務において多言語化された文書が蓄積されており，そのような文書対から対訳を抽出し，翻訳メモリの構築に役立てたいという要望がある．

ビジネス文書のなかには，プレゼンテーション形式のようにコンテンツが平面上にレイアウトされているものが多く存在している．このような形式の文書の場合，コンテンツの順序が自明でないため，順序が定まっていることを前提とした動的計画法などの既存手法による対訳文抽出を適用することができない．そこで本稿では，対訳関係の推定において，文書中のコンテンツを一次元に整序することなく，レイアウト，すなわちテキストを含むコンテンツの位置の情報を用いて，プレゼンテーション文書から高精度で対訳文抽出を行う手法を提案する．

2　関連研究

対訳文抽出における既存手法として，対訳辞書と動的計画法を利用してテキストの類似性と順序から対訳関係を推定する方法 [1] が存在する．また文書の構造を利用して対訳文抽出を行う手法として，HTML 構造を用いた手法 [2, 3] がある．

また，レイアウトを利用した情報抽出には，Web ページをブラウザでレンダリングした結果とヒューリスティクスを利用してノイズを除去する手法 [4] などが存在する．また，HTML やスプレッドシートに含まれる二次元の表の意味解析や情報抽出が盛んに研究されている [7, 8, 9]．

一方，プレゼンテーション形式の文書を扱う既存手法には，文書中に含まれるキーワードを利用した，プレゼンテーション蓄積検索システム [6] などがあるが，位置座標やコンテンツの種類を利用した情報抽出に関する研究は，まだあまり行われていないのが現状である．

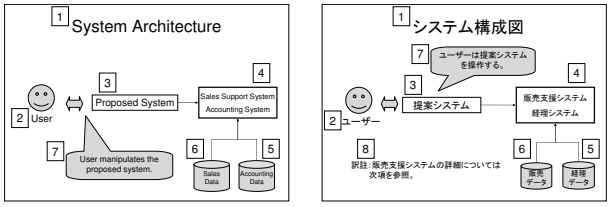

図 1: 対訳関係にあるページの例．左側が翻訳前の英語，右側が翻訳後の日本語のページである．1〜8 は，翻訳前のページの内部構造の順序に従って，両言語間の対応関係を示すために付けた便宜上のものである．

3　レイアウト構造を持つ既訳文書対からの対訳抽出

3.1　タスクの定義

本研究では，対訳関係にあるプレゼンテーション文書のページの対から，対応するテキストの単位である「オブジェクト」を同定する処理を実現する．なお，ページ相互の対応は予め取られているものとする．

図 1 に対訳のページ対の例を示す．多くの場合は対訳関係にあるオブジェクト同士の対応が取れているものの，3.2 節に示すように，追加，削除，移動等が行われることもある．

プレゼンテーション文書の場合，オブジェクトの中のテキストは短めの文やフレーズである場合が多く，場合によっては一つの単語の場合もあるが，一つのオブジェクトに複数文が含まれていることもある．翻訳メモリのリソースを構築する場合には文単位の対応に変換する必要があるが，オブジェクトの対応を段落の対応と見なせば，そこから文間の対応を取るのは既存手法により実現できる．

3.2　レイアウトの複雑さ

プレゼンテーション文書を翻訳する際には，WYSIWYG の環境で見栄えを重視して作業される傾向がある．その中でも，以下のような操作が，対訳関係の抽出の障害となっている．

表 1: 図 1 中のオブジェクトの順序．どの方法においても両言語間で一致しない．

内部構造	英語	1	2	3	4	5	6	7	
	日本語	1	2	4	6	7	5	3	8
x 座標順	英語	2	7	1	3	6	4	5	
	日本語	2	8	1	3	7	4	6	5
y 座標順	英語	1	4	3	2	7	6	5	
	日本語	1	7	4	3	2	8	5	6

- 翻訳作業に伴う編集の際に，例えばコピー＆ペーストで別個のオブジェクトを生成する処理や，アニメーションの調整などにより，オブジェクトの順序などの文書の内部構造を変化させる（例：図 1 中 5 ）．
- 翻訳後の文書に註を付けるなど，新しいオブジェクトを追加する（例：図 1 中 8 ）．
- 新しいオブジェクトの追加に伴って，スペースの都合などでオブジェクトを移動させる（例：図 1 中 7 ）．

二言語のアラインメントの既存手法を適用するには，文書中のオブジェクトを一次元の順序で並び替える変換が必要となるが，上記で見たような操作によって，翻訳の前後で順序が大きく変わってしまう．表 1 は，図 1 のオブジェクトを内部構造の順番[1]，x 座標，y 座標で整序した時の順序を示したものである．このように，両言語間の対応を取りやすくできるような一般的な並び替え方は存在しないため，対訳関係の抽出は自明でない．そこで，次節において，オブジェクトの移動や追加に対して頑健な手法を提案する．

4 提案手法

ここでは 3 節で述べたように文書中のコンテンツが平面上にレイアウトされた文書から対訳関係の推定を行うための手法を提案する．

4.1 2 部グラフマッチングへの帰着

二言語の文書から対訳を抽出する問題は，オブジェクトを頂点，対訳関係を枝，そのときのコストを重みとする 2 部グラフの最小重み最大マッチングを求める問題とみなすことができる．

このとき，3 節で述べたように，翻訳作業の前後でオブジェクトの数が変わる場合があるため，ダミーの頂点を作成して両側で頂点の数を揃えた．これにより，Hungarian 法などを用いて容易に解を求めることが可能になる．ある頂点とダミー頂点との間に枝が張られた場合は，その頂点に対応するオブジェクトは対訳関係を持たないことを意味し，枝のコストは $NoMatchPenalty$ を用いる．本研究では，問題を一般化するために作成するダミー頂点の数は反対側の頂点の数とする．つまり，翻訳前文書のオブジェクト数が n で，翻訳後文書のオブジェクト数が m であった場合は，両側にそれぞれ m,n 個のダミー頂点を作成し，両側の頂点数を $n+m$ 個に揃える．

翻訳前 s と翻訳後 t を結ぶ枝のコストを計算するにあたって，テキスト情報によって求められる $ContentCost$ とレイアウト情報から求められる $LayoutCost$ の 2 種類を用いる．$Cost(s,\phi)$ はオブジェクト s と対訳関係にあるオブジェクトが存在しない場合のコストを表している．

$$\begin{aligned} Cost(s,t) &= \alpha ContentCost(s,t) \\ &\quad +(1-\alpha)LayoutCost(s,t) \\ Cost(s,\phi) &= Cost(\phi,t) \\ &= NoMatchPenalty \end{aligned}$$

4.2 コンテンツ情報に基づくコスト

コンテンツ情報に基づくコストでは，コンテンツに含まれるテキストの長さと内容の一致度を用いて次式で定義される．

$$\begin{aligned} ContentCost(s,t) &= \beta TextLengthCost(s,t) \\ &\quad +(1-\beta)WordMatchCost(s,t) \end{aligned}$$

テキストの長さによるコスト

両言語のテキストに含まれる内容語 (名詞及び用言) の数の差を用いてテキストの長さがどれだけ離れているかのコストを次式により求める．

$$TextLengthCost(s,t) = \frac{2|N_s - N_t|}{N_s + N_t}$$

テキストの内容によるコスト

対訳辞書[2]を用いて辞書引きを行った結果が一致した語の割合を用いてテキストの内容がどれだけ乖離しているかのコストを次式により求める．また，二言語間で同じ英単語や数字が出現したときも対訳辞書でマッチした場合と同じ扱いとする．

$$WordMatchCost(s,t) = 2\left(\frac{Dict_{st} + Num_{st} + Eng_{st}}{N_s + N_t}\right)$$

[1] オブジェクトの前後関係を表す順番にも用いられ，新たなオブジェクトは位置に依らずに末尾に追加される．

[2] 実験では，名詞 130,306 語，用言 34,218 語のエントリを持つ日本語→英語の翻訳辞書を用いた．

ここで $Dict_{st}$ は対訳辞書を引くことにより対応する語が見つかった数，Num_{st} は同じ数字が両言語で出現した数，Eng_{st} は同じ英単語が両言語で出現した数を示している．また N_s,N_t は言語 s,t のテキストに含まれている内容語の数を示している．

4.3 レイアウト情報に基づくコスト

レイアウト情報に基づくコストでは，以下に述べるように二言語の文書間でのコンテンツ同士の座標距離と，コンテンツ同士の重なりの情報を用いて次式で定義する．

$$\begin{aligned} LayoutCost(s,t) &= \gamma EuclideanCost(s,t) \\ &\quad +(1-\gamma)OverlapCost(s,t) \end{aligned}$$

コンテンツ間の座標距離に基づくコスト

二言語の文書間で，コンテンツの左頂点の座標 (x,y) の距離を用いて次式でコストを求める．

$$EuclideanCost(s,t) = \frac{\sqrt{(s_x-t_x)^2+(s_y-t_y)^2}}{MaxDistance}$$

ここで $MaxDistance$ はその文書でとりうるコンテンツ間の最大距離を示す．本研究では対象としたプレゼンテーション形式の文書の縦幅，横幅から対角線の距離を求め，$MaxDistance$ として用いた．

コンテンツ同士の重なりに基づくコスト

二言語間のコンテンツを同じ空間に写像したときにコンテンツ同士の重なりがどの程度存在するかのコストを次式により求める．

$$\begin{aligned} OverlapCost(s,t) &= \eta OverlapWidthCost(s,t) \\ &\quad +(1-\eta)OverlapHeightCost(s,t) \end{aligned}$$

$$\begin{aligned} OverlapWidthCost(s,t) &= 1-\frac{OverlapWidth_{st}}{TotalWidth_{st}} \\ OverlapHeightCost(s,t) &= 1-\frac{OverlapHeight_{st}}{TotalHeight_{st}} \end{aligned}$$

このとき，図 2 に示されるように，$OverlapWidth_{st}$，$TotalWidth_{st}$，$OverlapHeight_{st}$，$TotalHeight_{st}$ はコンテンツ s とコンテンツ t の重なりから求める．

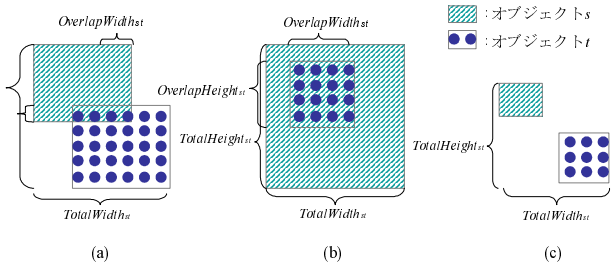

図 2: コンテンツ同士の重なりに関する値を求める例．例 (c) の場合は重なりが存在していないため，$OverlapWidth$，$OverlapHeight$ はともに 0 となる．

5 評価実験

5.1 実験の設定

英語から日本語に翻訳された各 112 ページからなる技術文書を用いて，対訳関係にあるオブジェクトの自動抽出を行った．提案手法のほか，対照実験として DP マッチングを試みた．

DP マッチング 以下の方法により一次元への並び替えを行ってから，$ContentCost$ のみを用いた場合と $ContentCost$ と $LayoutCost$ の両方を用いた場合についてそれぞれ動的計画法を適用した．

1. 内部構造順
2. x 座標順
3. y 座標順

2 部グラフマッチング レイアウトとコンテンツの情報を取り入れた柔軟なマッチングを実現するための提案手法である．以下の情報の利用を試みた．

- コンテンツ情報
- レイアウト・コンテンツ情報の組み合わせ

いずれの手法においても，4 節で定義したコストを用い，経験的に α, β, γ, η はすべて 0.5 に，$NoMatchPenalty$ は 1.0 に設定した．

5.2 実験結果

各手法により抽出された対訳関係から，それぞれ無作為に 100 対ずつを取り出して，それらが正しい対訳となっているかどうかを評価した．表 2 に，$ContentCost$ のみを用いた場合と $ContentCost$ と $LayoutCost$ の両方を用いた場合の結果を示す．

表 2 に示される結果のとおり，全ての手法において $ContentCost$ のみを用いた場合より $ContentCost$ と $LayoutCost$ を利用した場合の方が正解率は高いという

表 2: *ContentCost* のみを用いた場合と *ContentCost* と *LayoutCost* の両方を用いて対訳オブジェクト抽出を行った場合の正解率.

マッチング手法		正解率	
		Content	*Content* +*Layout*
整序→ DP マッチング	内部構造	69%	79%
	x 座標順	64%	65%
	y 座標順	69%	73%
2 部グラフマッチング		66%	95%

表 3: *ContentCost* と *LayoutCost* の両方を用いて対訳オブジェクト抽出を行った場合の相対的な再現率.

		抽出対訳数	相対再現率
整序→ DP マッチング	内部構造	1700	1
	x 座標順	1787	0.86
	y 座標順	1787	0.97
2 部グラフマッチング		1811	1.28

結果になった．特に 2 部グラフマッチングを用いた手法では，95% と非常に高い正解率を示した．

次に *ContentCost* と *LayoutCost* の両方を用いた場合の各手法により抽出される対訳関係の相対的な再現率を調べ，その結果を表 3 に示す．相対再現率は内部構造を用いた場合を基準とした．正しい対訳関係が得られる数の相対的な割合，すなわち，文書全体から抽出された対訳の数の比と，適合率の比を乗じたものである．表 3 にみられるように，2 部グラフマッチングを用いた手法は最も高い再現率を示し，他の手法と比べて適合率，再現率の両方において優れた結果となった．

5.3 考察

ContentCost のみの場合には「クラウドソーシング (日): crowdsourcing(英)」といった辞書に載っていない新しい技術用語や,「サービス開始 (日): introduction(英)」といった内容を分かりやすくするために意訳されたテキストではうまく対訳の推定を行うことができておらず，特に 2 部グラフマッチングを用いたときはその傾向が顕著であった．

ContentCost と *LayoutCost* を利用した場合においても，DP マッチングを利用した場合，オブジェクトの並び替え後の順序が二言語間で異なっている部分を正しく抽出できなかったケースが多く見られた．一方で，提案手法では 2 部グラフマッチングを用いたことで，オブジェクトの順序の変更に拘わらず多くの場合で正しく抽出することができていた．

また，*ContentCost* を利用した DP マッチングでは抽出できない対訳が多かったのに対して，提案手法ではレイアウト情報が有効に働いて正しい対訳を抽出することができていた．

但し，対訳辞書があたらない語を多く含む文で，かつ，二言語間のコンテンツの座標が重なりを持たないくらい離れている場合などは，提案手法においても正確な抽出を行うことができていなかった．このような対訳を抽出するには，構文情報やオブジェクトの形状などを考慮する必要がある．

6 まとめ

一次元の順序を持たないようなレイアウトを持つ対訳文書からも，高い精度で両言語の対応付けが行えることが実証できた．本研究は一般性を重視してオブジェクトの種類などは用いなかったが，文書の種類に特化した情報を取り入れることによってさらに精度を上げることが期待できる．また，多様な種類の文書から抽出された対訳関係を用いて，翻訳メモリを用いた翻訳業務の改善と，それに付随するインタフェース等の研究がさらに進展するであろう．

参考文献

[1] 淺利俊介, 竹内孔一, 阿辺川武, 影浦峡. “Web 上の兄弟ページを利用した対訳文書からの段落アラインメント,” 言語処理学会第 15 回年次大会発表論文集, 2009.

[2] P Resnik, NA Smith. “The web as a parallel corpus,” Computational Linguistics, Vol. 29, No. 3, pp.349-380, 2003.

[3] L Shi, C Niu, M Zhou, J Gao. “A DOM tree alignment model for mining parallel data from the web,” Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 489–496, Sydney, July 2006.

[4] 鶴田 雅信, 増山 繁. “未知のサイトに含まれる Web ページからの主要部分抽出手法,” 言語処理学会第 14 回年次大会発表論文集, 2008.

[5] 田仲 正弘, 石田 亨. “表構造の一般化に基づくオントロジの獲得,” 情報処理学会論文誌, Vol. 47, No. 5, pp.1530-1537, May 2006.

[6] 岡本 拓明, 小林 隆志, 横田 治夫. “プレゼンテーション蓄積検索システムにおける適合度計算の改善,” 電子情報通信学会第 15 回データ工学ワークショップ, 2004.

[7] Matthew Hurst. “Classifying table elements in html.” *In Proceedings of the 11th International World Wide Web Conference*, Hawaii, USA, 2002.

[8] Dekai Wu and Ken Wing Kuen Lee. “A grammatical approach to understanding textual tables using two-dimensional SCFGs.” *In 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*. Sydney, Australia: Jul 2006.

[9] Minoru Yoshida and Hiroshi Nakagawa. “Web Document Parsing: A New Approach to Modeling Layout-Language Relations.” *In Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR2007)*, pp.203-207, Curitiba, Brazil, September 2007.