2U-01

ドキュメント構造の自動抽出

鵜飼 理央†　　山本 博史‡

近畿大学大学院　　近畿大学

1. 序論

文に対して，構文解析や係り受け解析を行うことは，文中の各形態素の文法的，あるいは意味的な役割を知る上で不可欠である．同様に，ドキュメントの構造を解析することで，文あるいはその文中の形態素のドキュメント内の意味的役割を知ることができると考えられる．そして，この情報はキーワードの使われ方に依存した検索や，複数ドキュメントの自動要約に有用であると考えられる．

ドキュメントの構造抽出は[1]〜[4]をはじめとして種々試みられているが，本稿では統計量に基づいてドキュメント構造を自動的に抽出することを試みる．

2. ドキュメント構造の自動抽出

・ドキュメント構造の表現方法

ドキュメントにある構造が現れた場合，それに引き続き現れる構造には制約がかかる．本稿では，この構造間の制約関係を図１に示す逐次、並列、省略の３つの関係で表現し、ドキュメント全体の構造は，それらの組み合わせで表現されるネットワークで表現されるものと仮定する．

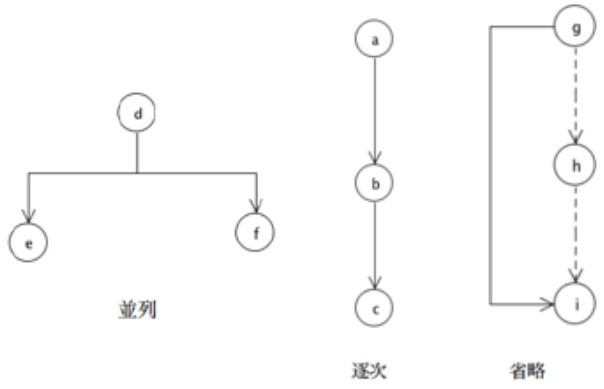

図 1: ネットワークの構造

・ドキュメントの構造決定

ドキュメントの構造の抽出のためには，各ドキュメントが取り得る構造に，ある程度の制約をあたえておくこととする．そこで，本稿では一定数のドキュメントをあらかじめ用意し，それに対して人手で構造を抽出することとする．そして、得られた複数の構造を全て包含するようなネットワークを生成し．他の全てのドキュメント構造はこのネットワークで表現可能であると仮定する．

・文のネットワークへの割り当て

次に，構造を抽出する対象のドキュメント中の各文を前節で得られたネットワークのノードに割り当てていく．この時に，同一ノードにはなるべく似かよった文が割り当てられることが好ましいと考えられるため．各ノードに割り当てられた文のノード内のエントロピーの和が最小になるように割り当てを行う．アルゴリズムとしては，ビタビとバウム＝ウェルチを用いる．

割り当ての手順は，以下のようになる．

1. ドキュメント構造の縦方向の分岐数を N とした場合．ドキュメント中の文を N 等分し，上から順番にノードを割り当てていく．横分岐に対してはドキュメントごとにランダムにどちらかに割り当てる．
2. 各ノードに割り当てられた文を用いて，ノードごとに文の生起確率を求める．これにより文がそれぞれのノードに割り当てられる確率を求めることができるようになる．
3. 次に 2)で得られた生起確率に基づき，各文をネットワークが与える制約を満たした上で，最も確率が高くなるノードに割り当てる．この時，ネットワーク上の複数の経路から同一ノードに到達する場合もあり得るが，その場合はビタビ最大値で近似するが，バウム＝ウェルチでは和を用いる．
4. 各ノードに含まれる文が更新されたため，2.) に戻って，生起確率を再計算する．
5. 2.) から 4.) までをエントロピー（パープレキシティ）が減少しなくなるまで繰り返す．

3. 実験

・実験条件

実験で用いた学習データは， "CiNii" [5]から収集した英語論文 10,203 件，日本語論文 30,002 件の abstract である．また，このうち 8,787 件は日英両

Automatic extraction of document structure

† Rio Ukai: Kindai University Graduate School

‡ Hirofumi Yamamoto: Kindai University

言語で同一内容が記載されている．また，各 abstract に含まれる平均文数は，日本語で約 5.7 文，英語で約 8.0 文であった．そして，この実験で論文の abstract を用いた理由は，それらは構造的に書かれていると考えられるからである．
収集した abstract を文単位，形態素単位に分解する．日本語の形態素解析には ChaSen[6] を用いた．また，形態素の生起確率は 1-gram 言語モデルを用い，学習には srilm[7]を用いた．

・実験結果
学習データからランダムに抜き出した日本語 100 ドキュメントから人手で構造を抽出した．その結果，図 2 に示すような 6 状態からなるネットワーク構造が得られた．

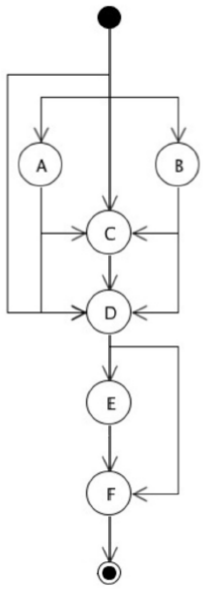

図 2: 人手で得られたネットワーク構造

次に，日本語，英語の abstract を 2 章の手順に従って各ノードに割り当てを行った． その結果日本語では初期状態のパープレキシティが 504 であったのに対し最終的に 445 に，英語では 940 から 798 に収束した．

・評価
評価に関しては，以下の 2 点を比較対象とし，本手法で得られた日本語 abstract の構造の境目と人手で得た境界との一致度で評価を行う．ただし，人手の境界は揺れを考慮して，二人の評価者を用意し，適合率では両者の OR を，再現率では AND を用いて評価した．二人の評価者の一致度の F 値は 88.0%であった．
それぞれの比較対象に対する適合率，再現率，F 値の結果を表 1 に示す．この結果から，人手で抽出した構造に対する一致度は 7 割前後と十分とは言えない結果になった．そして，手法による一致度に大差ないことが分かった．

表 1: 適合率・再現率・F 値

	適合率	再現率	F 値(F)
ビタビ	70.7%	63.9%	67.1%
バウム＝ウェルチ	72.3%	62.0%	66.8%

4. 結論と今後の課題

本手法により，ドキュメントを自動で構造ごとに分割することを可能となった．得られた構造は，必ずしも人間の感覚とは合わないが，同一内容の日本語のドキュメントに対しては，ほぼ同一の結果が得られたため，首尾一貫した構造抽出は行うことができたと考えられる．
今後の課題としては，構造の境目の一致度をさらに上げる必要があることである．また，抽出自体を自動で行っていく必要がある．

参考文献

[1] 土井美和子，福井美佳，山口浩司，竹林洋一，岩井勇(1993) 文書構造抽出技法の開発
[2] 黒橋禎夫，長尾眞(1994) 表層表現中の情報に基づく文章構造の自動抽出
[3] 甲斐郷子，中村順一，吉田將(1995) 表層表現に基づく文章構造解析を利用した論文改訂システムの試作と評価
[4] 前田浩邦，山肩洋子，森信介 (2014) 検索・分析のための手順文章からの意味構造抽出
[5] CiNii http://ci.nii.ac.jp/
[6] ChaSen -- 形態素解析器 http://chasen-legacy.osdn.jp/
[7] SRILM - The SRI Language Modeling Toolkit http://www.speech.sri.com/projects/srilm/