【予稿集】

機械による情報の縮約可能性について

松田裕幸*，天野晃**
*東京大学大学院学際情報学環学府，**国立情報学研究所
*matsudayuko@acm.org, **amano@nii.ac.jp

構文構造は意味空間が構成する意味の一部を隠れパラメータとして保持できるとして，たとえば潜在的意味論の理論が発達してきた．そうした理論の集大成の 1 つとして，LDA(Laten Dirichlet Allocation)が知られている．一方，グラフ理論も構文要素間の関係をグラフとして表現することで，関係意味を隠れパラメータとして浮かび上がらせることに成功している．本研究では，LDA にグラフ理論を組み合わせることで，大規模文書の縮約を試みる．

On Machine Abridgement of Documents

Yuko MATSUDA*, Kou AMANO**
*Graduate School of Interdisciplinary Information Studies, the University of Tokyo
**National Institute of Informatics

1. はじめに

大量の文書を前にしたとき，その縮約版を手動で作成することは容易でない．本研究では，機械による自動縮約を試みる．その際，教師データを期待できない環境において，それなりの品質を保ったまま高速に縮約版を生成するにはどのようにしたらよいだろうか. ここでは, まず LDA[1]によって文書内トピックを決定する．ついでトピックを hub とし，hub に接続する形で各文書を頂点接続したグラフを形成する．最後にトピックを頼りに出発文書と到着文書間（出発と到着は目的に応じて，その都度指定する）の半順序を求め，トピック間に所属する文書をさらに要約，列挙することで全体の縮約を完成させる．文書群の性質によってはトピック抽出に高い精度が必ずしも得られないことも多いが，トピックを hub に持ち，各文書を hub に接続させたグラフの形で文書群を構造化することで，本来隠れていたプロットを曖昧な形であれ明示的に浮かび上がらすことができ，それによって半順序関係の計算が可能となり，縮約版候補を生成することができる．当然，半順序が生成するパスには精度の高いものもあれば低いものが発生する可能性があるが，大量の文書を前提としたとき，同じ情報が複数の形で繰り返し登場することを想定する．

本研究では，トピック（主題）が曖昧になりがちな小説（『変身』英語版[2]）を敢えて選び，縮約を試みる．全体は 101 の節からなり，各節を文書とみなし，小説全体を文書群とし，トピック解析およびトピックに基づいたグラフ化を行い，縮約文を生成する．なお形態素として名詞，動詞を選択する（表 1)。

表 1. 原文(doc)と形態素(tokens)

	doc
0	One morning, when Gregor Samsa woke from trou...
1	“What’s happened to me?” he thought. It wasn’t...
2	Gregor then turned to look out the window at t...

tokens
[morning, Gregor, Samsa, wake, dream, find, tr...
[happen, think, dream, room, room, lie, wall, ...
[Gregor, turn, look, window, weather, drop, ra...

2. 実験結果

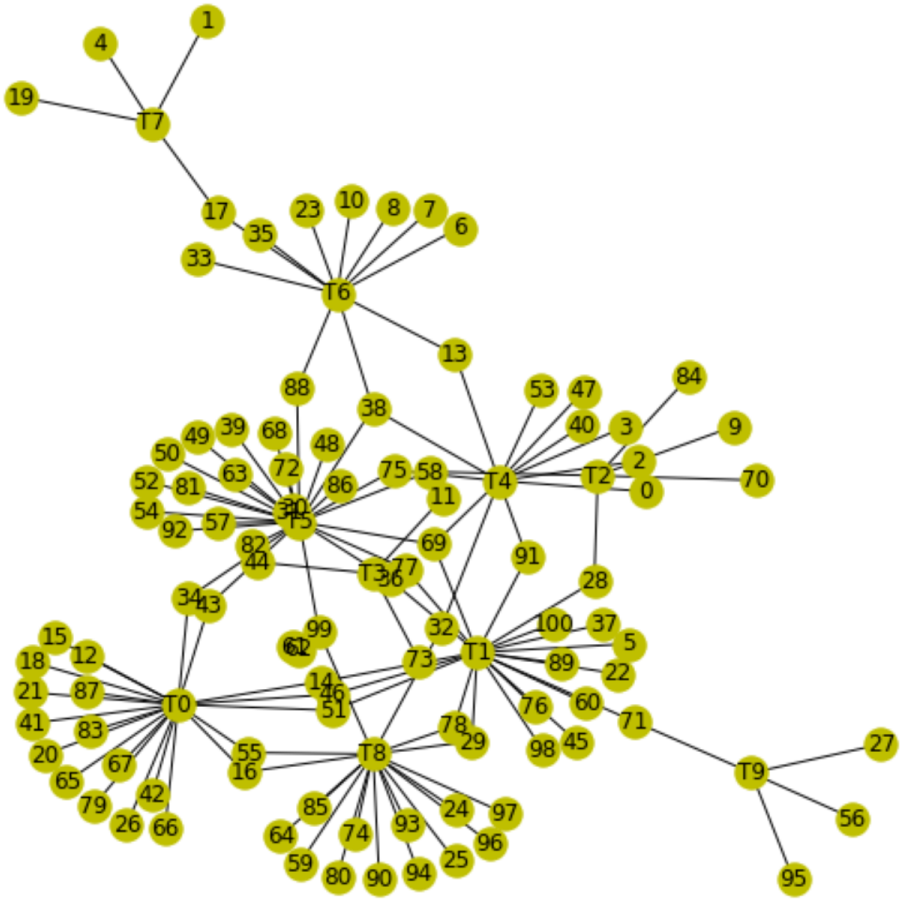

図 1. 文書-トピック関係グラフ

図 1 で 0 から 100 までの数字は各文書に対応する頂点を表し，T で始まる名前の頂点はトピックに対応する．トピック頂点から各文書に接続する辺はそれらの文書が共通するトピックを持つことを表す．本稿作成時では，半順序生成アルゴリズムは完成しておらず，出発文書 0 番，到着文書 100 番および, 共通トピック T4 および T1 を両端に持つ文書 69 番と 91 番に対し，0 番に属する全文ならびに残りの文書から共通トピックを「含まない」文を抽出したものを暫定の縮約版とする．共通トピックを「含まない」としたのは，共通トピックを含まない部分に潜在トピック外の意味があると仮定したためである．

【考察】上記トピック-文書関係グラフの構造は，形態素解析辞書[3]の選択，ならびにトピック解析に必要な辞書の作成とコーパス量およびトピック数に大きく依存する[4]。なお，グラフ理論および実装には networkX[5]を利用した。

3. 参考文献

3.1 単行書

[1] トピックモデルによる統計的潜在意味解析：佐藤一誠．コロナ社，2015.

3.2 ウェブサイト

[2] Project Gutenberg. https://www.gutenberg.org/

[3] spaCy. https://spacy.io/

[4] GENSIM. https://radimrehurek.com/gensim/

[5] networkX. https://networkx.org