3D4-1

論文における要約記述に対応するパラグラフの同定手法

Finding Corresponding Paragraph with Abstract Sentence in Academic Papers

亀田 尭宙[*1] 李 元[*2] 内山 清子[*1] 武田 英明[*1] 相澤 彰子[*1*2]
KAMEDA Akihiro LI Yuan UCHIYAMA Kiyoko TAKEDA Hideaki AIZAWA Akiko

[*1]国立情報学研究所 National Institute of Infomatics
[*2]東京大学 The University of Tokyo

Abstracts of scientific papers play a crucial role for efficient access to scientific literature. However, correspondance between abstract sentences and paragraphs in body text isn't given to researchers. We proposed two methods for finding corresponding paragraph with abstract sentence in academic papers. One is based on Longest Common Subsequence, which can capture apparent similarity, and the other is based on Latent Dirichlet Allocation, which can capture latent topic similarity. These methods will enable us to retrieve, understand and structurize knowledge more efficiently through implementation to digiral libraries.

1. はじめに

自分の研究や調査に役立つ学術論文を探す際に，論文のアブストラクトは重要な役割を担っている．アブストラクトには論文の目的，手法，貢献といった重要な情報が簡潔に書かれているため，その論文の本文を読む必要があるかどうかの判断を下す手掛かりとして有用である．しかしながら，その簡潔な記述は詳細さを犠牲にしているため，アブストラクトだけでは分からない部分について本文中の詳細な記述を得たいといった要求はしばしば生じる．このような要求にこたえるために，本論文では，アブストラクト中の文と本文中のパラグラフとの対応を発見する方法を提案する．提案手法をデジタルライブラリ中の検索システム実装することで，より効率的な学術情報探索が可能になると考えている．以下，2. 章でアブストラクト中の文と本文中のパラグラフとの対応を人手で分析した結果について述べ，3. 章でその分析に基づいた手法を提案し，実験によってその手法の効果を確かめた．

2. データのアノテーションと分析

我々は ACL Anthology Reference Corpus (ACL ARC) [Bird 08] から 30 の論文を選び，アブストラクト文と論文本文中のパラグラフとの対応付けを人手でアノテーションした．一つのアブストラクト文が複数のパラグラフに対応することも，逆に複数のアブストラクト文が一つのパラグラフに対応することもあるため，対応は多対多の関係になっている．表 1 にデータセットの統計を示す．

また，対応の種類や対応を発見する手掛かりについてもアノテーションを行った．このアノテーションは後述の実験において，エラー分析に用いた．それぞれの関連は表 2 のうち上の 3 つのいずれかのタイプが割り振られており，また別途，実験には用いなかった情報である章や節のタイトルが対応の手掛かりになり得たどうかについてラベルが付与されている．よって，上 3 つのタイプの合計数 217 は関係の総数と一致する．「同じ表現が出現する」というのは語のレベルではなく，フレーズのレベルで，専門用語や表現の使いまわしといったものが存在したことを指している．

連絡先: 亀田 尭宙, 国立情報学研究所, 東京都千代田区一ツ橋 2-1-2, kameda@nii.ac.jp

表 1: データセットの統計

論文の総数	30
1 論文あたり平均のアブストラクトの文数	4.9
アブストラクトの 1 文あたりの平均の関係数	1.5
1 論文あたり平均のパラグラフ数	43.5
内アブストラクトと対応をもっている数	5.6

表 2: 関連タイプの統計

同じ表現がほぼ出現しないが 意味において関連があるもの	130
同じ表現が一部出現するが 文の構造が異なるもの	48
同じ表現が連続して出現し それによって明白に関連が分かるもの	39
章や節のタイトルに 対応する語が出現するもの	36

3. 提案手法

3.1 最長共通部分列に基づく手法

我々は，あるアブストラクト文について最長共通部分列 (Longest Common Subsequence, LCS) の探索を元に各パラグラフ文に対する類似度を計算し，その類似度が最も高かったパラグラフを対応パラグラフとする手法を提案した [李 12]. これは，アブストラクト文の多くにおいて，論文本文と同じ言い回しがそのまま用いられることが多かったためである．類似度の計算手法は，以下のとおりである．まず語 t に対してその重要度を示す重み w_t を t がデータセット全体に現れる頻度 f_t を用いて以下のように定義する．

$$w_t = \frac{1}{log(f_t) + 1} \tag{1}$$

これを用いてアブストラクト文 i とパラグラフ文 j の最長共通部分列 $lcs(i,j)$ の重みを $\sum_{t \in lcs(i,j)} w_t$ と算出する．一方で，アブストラクト文 a とパラグラフ文 p の一致度を，アブストラクト文の語数 $|a|$, パラグラフの語数 $|p|$, 最長共通部分列の語数 $|lcs(a,p)|$ によって，$\frac{|lcs(i,j)|^2}{|a| \cdot |p|}$ と算出する．これ

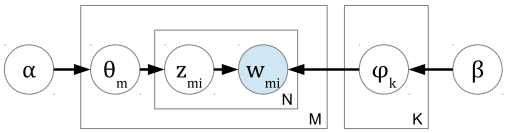

図 1: LDA のグラフィカルモデル

らを組み合わせて

$$\frac{|lcs(i,j)|^2}{|a|\cdot|p|}\sum_{t\in lcs(a,p)} w_t \tag{2}$$

を類似度として用いた.

3.2 トピック分布の類似度に基づく手法

我々は， Blei らの提案した Latent Dirichlet Allocation (LDA) [Blei 03] というトピックモデルを用いて，語やその集合としての文章の背後のトピック分布の類似度に基づいて，アブストラクト文ともっともトピック分布の近い本文中のパラグラフを対応パラグラフとする手法を提案する．表現の使いまわしが多いとはいえ，表 2 の統計を見る限り，表現が異なるために最長共通部分列に基づく手法では捉えることの難しそうな関係も多い．潜在的なトピックの類似度という一段階抽象的な意味での類似度を測ることによってそういった関係を捉えようと考えた.

LDA のグラフィカルモデルは図 1 のようになっており，α をパラメータとするディリクレ分布に従って生成された M 個のドキュメントそれぞれの多項トピック分布から，語に割り当てられる N 個のトピック $z_{mi}(i=1...N)$ が選択され，そこから語 w_{mi} が生成される．一方で，決められたトピックから語を生み出す生成モデルは β をパラメータとするディリクレ分布に従って生成された K 個の多項分布 ϕ_k が担う．今回のタスクにおいては，一般的な LDA の利用法より一つ粒度を細かくし，文や段落を「文書」とみなし，その集合である各論文をモデル化する．トピック割り当ての推定には Collapsed Gibbs sampling を用い [Carpenter 10]，分布の推定はトピック割り当ての推定に基づいて以下の式で行った [Griffiths 04].

$$\hat{\theta}_{m,k} \quad = \quad \frac{n_{m,k}+\alpha}{\sum_k n_{m,,k}+\alpha} \tag{3}$$

文書 m 内に出現する，トピック k を割り当てられた語の数を $n_{m,k}$ とし，推定したトピックの確率 $\hat{\theta}_{m,k}$ を文書ごとにベクトルにまとめたものが，文書のトピックベクトルになる．これをアブストラクト文と本文パラグラフについて算出し，それらのコサイン類似度を最終的な類似度とすることで，対応付けを発見した.

4. 実験結果と考察

実験には 2. 章で述べたデータセットを用い，人手のアノテーションを正解とし，精度と再現率を算出し，エラー分析を行った.

結果を表 3 に示す．LCS は最長共通部分列に基づく手法を，LDA はトピック分布の類似度に基づく手法をそれぞれ示している．全体的に LCS の方が良い結果を示した．これは，対応パラグラフにフレーズレベルでなくとも語のレベルでの一致は多かったこと，LDA の学習において外部コーパスを用いず一文書のみを用いたため，十分にモデル化できなかったトピックが存在したことが大きな要因と考えられる．一方で，エラー分析から，他のラベルと相対的に「同じ表現がほぼ出現しない」とアノテーションされた関係やタイトルで対応する語が存在する関係について LCS に比肩する結果を示しており，また，LCS では捉えられなかった関係についての正解もあったことから，学習過程を工夫することで表面上は一致が多くない関係を発見できる可能性がある.

表 3: 実験結果

	LCS	LDA
精度	45.3%	12.8%
再現率	30.9%	8.8%
同じ表現がほぼ出現しないが 意味において関連があるもの	70	79
同じ表現が一部出現するが 文の構造が異なるもの	10	28
同じ表現が連続して出現し それによって明白に関連が分かるもの	1	22
章や節のタイトルに 対応する語が出現するもの	14	16

5. おわりに

本論文では，効率的な学術情報検索に寄与するために，論文のアブストラクトと本文のパラグラフの対応を発見する手法を提案し，最長共通部分列に基づく手法とトピック分布の類似度に基づく手法の 2 つを比較した．前者が優れた性能を示し，前者で捉えきれない関係を後者が捉えられていることを示した．今後はトピック分布の類似度に基づく手法を改善するとともに，2 つの手法を組み合わせることで性能を改善する方法についても模索したいと考えている.

参考文献

[Bird 08] Bird, S., Dale, R., Dorr, B. J., Gibson, B. R., Joseph, M., Kan, M.-Y., Lee, D., Powley, B., Radev, D. R., and Tan, Y. F.: The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics, in *LREC* (2008)

[Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent dirichlet allocation, *The Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003)

[Carpenter 10] Carpenter, B.: Integrating out multinomial parameters in latent Dirichlet allocation and naive Bayes for collapsed Gibbs sampling, Technical report, LingPipe Inc. (2010)

[Griffiths 04] Griffiths, T. L. and Steyvers, M.: Finding scientific topics, *Proceedings of the National Academy of Sciences*, Vol. 101, No. Suppl. 1, pp. 5228–5235 (2004)

[李 12] 李 元, 亀田 尭宙, 内山 清子, 相澤 彰子 : A Method for Corresponding Paragraphs with Sentences in Academic Paper's Abstract, 第 74 回情報処理学会全国大会講演論文集 (2012)