トピック分析とtf-idfによる単語の専門性推定

川本 峻頌　　高木 友博
明治大学大学院理工学研究科情報科学専攻
{syunyo, takagi}@cs.meiji.ac.jp

1 はじめに

本研究の潜在的な目標は, ユーザごとにターゲティングした文生成モデルの構築である. 近年, 大規模データを利用することで, 深層ニューラルネットワークを用いた手法が高い性能を示している. さらに強化学習と組み合わせることで, BLEU や ROUGE 等の評価指標を報酬として直接最適化するモデルが注目されており, 高い性能を実現している. しかし, 報酬となる BLEU や ROUGE は参照文との一致度しかみておらず, 人による評価との相違が問題として挙げられている. また文生成における他の報酬として, 原文との意味の関連性や, 文の流暢性, 一貫性等の評価が提案されているが, これらはターゲティングした文生成モデルを目的とした評価値ではない. ユーザはプログラミングや政治等のトピックに対してそれぞれ専門性をカテゴリ分布のような形で持っており, 文も同様に専門性を持っている. ユーザの専門性に近づくように生成された文は, ユーザにとってわかりやすい文になっていると考えられる.

本研究では LDA によるトピック分析と tf-idf を用いた単語の専門性を推定する手法を提案し, その有効性を示す. 以下, 2 節では専門性に関する関連研究について述べる.3 節では提案手法の詳細, 4 節では評価実験, 5 節では考察について述べ, 6 節で本稿を締めくくる.

2 先行研究

2.1 文書の難易度

専門単語の専門性は文書の読みやすさに関わるため, 文書の難易度と関連する. 既存研究では, 教科書などの一般コーパスから文字種の頻度や文長を基準とし推定するもの [1][2] や, 難易度つきコーパスを用いて教師あり学習を行うもの [3] などがある. しかしこれらの難易度は一般コーパス全体に対して測るため, ユーザの嗜好を反映するものではない.

2.2 特定のトピックに対する単語重要度

専門性に関する既存研究として, 特定のトピックにおける単語の重要度を計算する手法が提案されている.

片山ら [5] はユーザの特定分野に関する知識の推定に LDA を用いたトピック適応を行なっている.[5] では, トピックにおける用語の頻度と親密度には高い相関関係があるとして, LDA での各トピックにおける単語の生起確率を, トピックにおける難易度としている.

滝川ら [4] は特定分野にどれだけ精通しているかを判断することを目的とした単語重要度計算手法として CrRv(Category relevance Rarity value) を提案している.CrRv では, 専門辞書には一般人も使用する単語（例えばプログラミングの場合, 「java」など）が含まれているのが一般的であり，専門辞書に含まれる単語の中でも一般人があまり用いない単語に高い重要度を付与する．つまり， 特定分野にどれだけ精通しているかを判断するために，該当分野に精通していないと知り得ない単語に高い重要度を与える.

3 提案手法

CrRv[4] では特定分野に関する大量の辞書を用意する必要がある. 滝川らは分野をプログラミングに絞り, Web サイトや用語集, 辞典などから辞書を構築していた. しかし, この場合, 特定の分野各々に対し辞書を構築する必要があり, 分野の選定や収集コストを要する. 我々は上記の問題を解決するために, 一般コーパスから, 分野の分割, 選定, 単語と分野の紐付けを自動化する手法を提案する. 提案手法では, 文書と専門用語とトピックの関係を以下として仮定する.

- 文書は潜在的にトピックを複数保有しているが, その中である一つのトピックに関して主張するものとして書かれている.
- 専門用語は, 一般的にも該当トピック内でもあまり使用されない単語である.
- 専門用語は, 他のトピックと比較して該当トピック内での出現頻度が高い単語である.

3.1 トピック分析による専門辞書構築

文書は潜在的にトピックを複数保有しているが, その中である一つのトピックに関して主張するものして書かれている, という仮定の下, 文書の主張するトピックを LDA により 1 つに特定する. 特定したトピックより, トピックごとに文書を集計し, 専門辞書を構築する.

3.2 専門性推定

提案手法を式 (3.2.1) に示す.

$$TrRv(t,p) = \frac{\frac{DF_P(t)}{|D_p|}}{\frac{DF_P(t)}{|D_p|} + \alpha * \frac{DF_N(t)}{|D_n|}} \quad (3.2.1)$$

$$\alpha = \frac{\sum_{t'}^{T} DF_P(t')/|D_p|}{\sum_{t'}^{T} DF_N(t')/|D_n|} \quad (3.2.2)$$

上式において, 単語を t, 対象とするトピック p の文書集合を D_p, 対象トピック以外の文書集合を D_n, 全文書集合を $D(= D_p + D_n)$ で表す.$TrRv$ はトピック p における単語 t の専門性を示す.全単語集合を T, D_p の文書数を $|D_p|$, 単語 t の D_pにおける文書出現頻度を $DF_p(t)$, 単語 t の D_nにおける文書出現頻度を $DF_N(t)$ としている.

式中, $TrRv(t,p)$ は単語 t のトピック p への出現頻度の偏り具合を示し, p への偏りが強い単語に大きな値をとる.$|D_p|$ と $|D_n|$ は同一ではないため正規化している.αは単語 t が p 以外のトピックに出現した際に重要度を下げる割合を調整するパラメータである.

3.3 専門性推定の流れ

単語の専門性を推定する全体の流れは以下の通りである.

1. 一般コーパスを収集する.
2. LDA によるトピック分析を行い, 最も所属確率の高いトピックを, 文書のトピックとする. トピックごとに文書を集計し専門辞書を構築する.
3. 各トピックにおける単語の DF 値と, そこに含まれる各文書における単語の TF 値を集計する.
4. 集計した DF, TF 値から TrRv を求め, これを専門性を表す評価とする.

4 評価実験

本節では評価実験で使用したデータの内容と, 実験概要, 評価方法について説明した後, 実験結果とその考察を述べる.

4.1 使用データ

実験では livedoor ニュースコーパスを用いた. 当コーパスは NHK Japan 株式会社が運営する「livedoor ニュース」のうち, 9 つのクリエイティブ・コモンズライセンスが適用されるニュース記事を収集し, 可能な限り HTML タグを取り除いて作成されたものである. 総記事数は 7,367 件である.

前処理として, 本文では 1 行目に該当記事 URL, 2 行目の投稿日付を除去した. さらに, 記事本文に含まれる URL 除去の上, 名詞のみを抽出し, 3 文字以上かつ DF 値が 30 以上の単語を使用した. 総単語数は 77,674, 処理後の単語数は 5,196 である.

4.2 評価方法

処理後の単語を 500 件サンプリングし, コーパスから想定される特定のトピックに関して「専門度」を 0-5 段階で作業者 5 名にてアノテーションを行なった. このアノテーションの平均値と各手法における専門性の評価値とのピアソン相関を計測し評価する. 有意水準は 5%とした. 比較手法として 2.2 で示した [5] を参考に LDA による生起確率を用いる.

表 4.1: トピック 1 における評価値のピアソン相関

手法	提案手法	LDA による生起確率
ピアソン相関	0.57	0.28

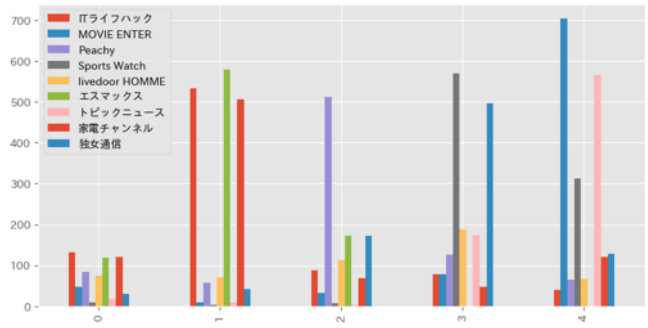

図 4.1: LDA でのトピック毎の文書頻度

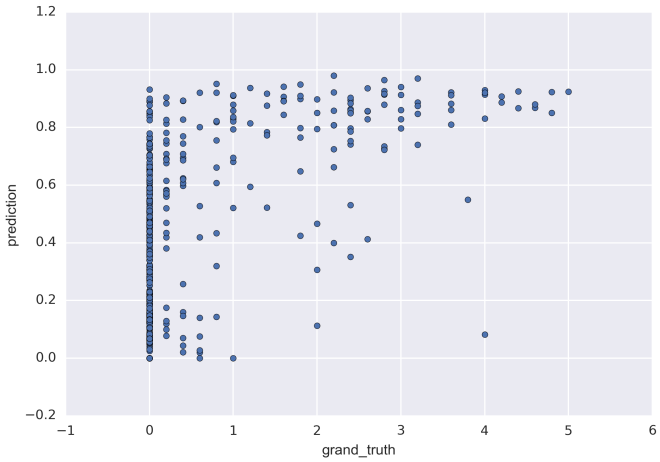

図 4.2: トピック 1 における提案手法と人の評価値

4.3 実験結果

LDA の潜在トピック数は 5 とした.LDA によるトピック分析の結果, トピックごとの記事の頻度分布は図 4.1 のようになった. トピック 1 では IT, 家電に関する記事が多く, トピック 2 では女性向けニュース, トピック 3 ではスポーツ, 芸能関連, トピック 4 では映画情報に関する記事が多い.

以降トピック 1 に着目する. トピック 1 における各手法での単語重要度と, アノテーション値とのピアソン相関値を表 4.1 に示す.

5 考察

表 4.1 の結果から, 提案手法の方が高い精度で専門性を推定ができていることがわかる. このことから, ある単語の特定のトピックにおける専門性は, トピックでの出現頻度に加えて, 出現頻度の低さや, 他トピックにおける出現頻度も考慮するべきであることがわかる.

図 4.2 は単語に対する提案手法による評価値と, 人の評価値との散布図である. 図 4.2 より, 人の評価値が大きくなるほど精度の高い推定ができているが, 全体的に提案手法は高い評価値を与えやすいことがわかる. 特に, 提案手法では高い評価値を与えているが, 人は低い評価値を与えている単語が多い. これらの単語例を表 4.2 に示す.

表 4.2: 提案手法で高く, 人は低く評価している単語例

Sleep, スロット, 最新情報, 世界初, 大きめ, 手書き, プッシュ, オススメ, 新生活, リーダー, 当たり, お祭り, ユニーク, スマート, 現地時間, プレミアム, ローカル, 株式会社, ホット, note,

表 4.2 中, 「Sleep, スロット, プッシュ, リーダー」などの単語では多義性による影響が確認できる. 例えば「スロット」は一般的にパチンコスロットの意味で使用されやすいが, IT, 家電トピックでは PC でのマザーボードを挿す穴の名称として用いられる. このように, 使い方によっては専門性が高くなるような多義語に, 提案手法は高い評価値を与えている. 一方で, これらの単語は一般的に専門性の低い意味の方が想起されやすいため, 人の評価値は低くなり, 提案手法とのずれが生じたと考えられる.

また, 「新生活, 当たり, お祭り, プレミアム」などの単語は, 意味としては専門性は低いが, IT, 家電トピック内の文書では「ハイテク家電で新生活」や「新生活に向けて」などの決まり文句が頻繁に使用されるため, 提案手法では高い評価値を与えている.

6 おわりに

本稿では, トピックにおける単語の専門性の推定を行なった. 既存の計算手法では, トピックの選定や, 特定トピックに対しての専門辞書の収集が必要であった. 本稿では, LDA によるトピック分析を行い自動で専門辞書を構築することで, 上記の問題を解決した. 提案手法をニュース記事における IT への専門性評価で実

験を行なった結果, 既存手法と比較して, 提案手法の方が高いピアソン相関を得ることができた.

一方で, 提案手法では多義語に対して上手く対応できていない. 今後, 文の専門性を推定する際には, 前後単語等の文脈情報を考慮するなど, 語義曖昧性について検討する必要がある. また, 該当トピック内で意味に関わらず, 決まり文句のように頻繁に用いられる単語にも適切な重みを与えることも今後の課題である.

本研究の潜在的な目標は, ユーザごとにターゲティングした文生成モデルの構築であるため, この後のステップとして, 提案手法を応用しユーザと文の専門性を評価し, 報酬設計を行う. その後, 強化学習を用いた文生成モデルへの組み込みを行い, 特定のユーザにとってわかりやすい文生成を目指す.

参考文献

[1] 近藤陽介, 松吉俊, 佐藤理史. 教科書コーパスを用いた日本語テキストの難易度推定. 言語処理学会第 14 回年次大会論文集, pp. 1113–1116, 2008.

[2] 小島健輔, 佐藤理史, 藤田篤. 文字 bigram モデルを用いた日本語テキストの難易度推定. 言語処理学会第 15 回年次大会論文集, pp. 897–900, 2009.

[3] 水谷勇介, 河原大輔, 黒橋禎夫. 日本語単語の難易度推定の試み. 言語処理学会第 24 回年次大会論文集, pp. 670–673, 2018.

[4] 滝川真弘, 山名早人. 和英短文を対象とした著者専門性推定への応用. 第 10 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2016), 2018.

[5] 片山太一, 小林のぞみ, 牧野俊朗, 松尾義博. トピック情報を利用したユーザの知識推定. 人工知能学会全国大会予稿集, 2013.