6P-2

論文における目的を含むパラグラフの推定

松田昇悟† 當間愛晃†
†琉球大学工学部情報工学科

1 はじめに

因果関係に関する知識は、ある物事についての理解を深めるためには非常に重要なものである [2]。これは自然言語処理においても同様であり、文書中に記述されている物事の理解を深めるためにはそれに関する因果関係を把握する必要がある。因果関係はある原因に対してある結果があるとする考え方であり、本稿では目的は原因に含まれると考えている。実際に因果関係抽出において目的を要素としているものもあるが、明確に目的として推定しているものは少ない。しかし、目的を推定するメリットはあり、文書における目的を示すことで、その文書で伝えようとしていることを把握しやすくなり、より深い文書理解につながると考えている。

本研究では、論文中に含まれる因果関係知識の一つである目的に着目し、目的を含むパラグラフを推定する手法について検討する。

2 関連研究

文章中における因果関係を扱った研究としては因果関係ネットワーク [1, 2] がある。これは複数の文書を対象に、それぞれを因果関係の要因のノードと結果のノードとして有向グラフを構築している。これらの研究では因果関係を抽出する際に「ため」のような手がかり語を用いて抽出している。特に澤村ら [1] は、手がかり語が表す関係を「理由」、「条件」、「目的」、「逆接」と表し明確にしている。

3 提案手法

本研究では論文における目的を含むパラグラフを推定する手法として、名詞と動詞を素性として用いる手法、手がかり語を素性として用いる手法、品詞と手がかり語の両方を素性として用いる手法を使用し、それぞれで目的を含むパラグラフを推定可能であるかを検討する。推定するものをパラグラフに選んだ理由としては、単文では単語数が十分ではないと判断したためである。今回推定しようとしている目的とは、論文における研究に関するものである。以下に目的の詳細と各手法について記述する。

研究目的
論文で記述されている研究の目的

実験目的
研究における実験の目的

それ以外
論文に記述されている上記以外のもの

3.1 品詞（名詞、動詞）を素性として用いる手法

目的を含むパラグラフの特徴が品詞に現れると仮定した手法である。品詞の中でも文章の内容を表している名詞と動詞を選択し、それ以外の品詞は考慮しない。得られたパラグラフを一つの要素として形態素解析を行い、単語が出現するかどうかをベクトルの要素とする。形態素解析には MeCab[4] を使用している。

3.2 手がかり語を素性として用いる手法

品詞を素性として用いる手法と同様に論文の各パラグラフを一つの要素として、手がかり語の素性が含まれているか含まれていないかを素性として用いる手法である。この手法で利用する素性は [1] で用いられているものに加えて、目的を含むパラグラフに頻出する表現や参考文献を示す際に記述される「[1]」なども「[数字]」として処理し、素性の一つとして扱う。さらに実験結果のような数値を一つの素性として扱うために「数字」として素性に加えた。ここでいう手がかり語とは、用いられている文節がどういう役割を持っているかを知るための手がかりとなる語のことであり、乾ら [3] では、「ため」という手がかり語について研究が行われている。今回使用する手がかり語の一覧を以下に示す。

Estimate a paragraph include purpose in the paper
Shogo Matsuda†, and Naruaki Toma†
†Department of Information Engineering, Faculty of Engineering, University of the Ryukyus

表 2: 評価値一覧

手法	研究目的			実験目的			それ以外		
	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値
品詞	0.63894106	0.58555556	0.60193683	0.4252381	0.28095238	0.3216589	0.84262979	0.88808081	0.86409171
手がかり語	0.67817599	0.70666667	0.68407805	0.	0.	0.	0.83116142	0.93090909	0.87778936
両方	0.65665584	0.60555556	0.62350175	0.38952381	0.26428571	0.29172605	0.84381673	0.88590909	0.8634891

表 1: 手がかり語一覧

「ため」、「ので」、「べく」、「本研究」、「本稿」「本手法」、「[数字]」、「数字」

3.3 品詞と手がかり語の両方を素性として用いる手法

上述の 2 つの手法で使っている素性 (名詞、動詞、手がかり語) の両方を用いる手法である。

4 評価実験

4.1 実験目的

本実験では、それぞれの手法でどの程度論文における目的を含むパラグラフを推定できるかを評価することが目的である。題材として論文を選択した理由は、より明確に文中に目的が書かれていると考えたためである。

4.2 実験概要

目的を含むパラグラフかどうかを識別する多値分類を行う。実験用データは、情報処理学会第 75 回全国大会の論文集からランダムに原稿 36 本を選択し、その中から得られた全 608 パラグラフに対して、「研究目的」、「実験目的」、「それ以外」として一人でタグ付けを行ったデータを用いた。それぞれの目的の数は、「研究目的」が 99、「実験目的」が 61、「それ以外」が 448 である。分類器には SVM を使用し、10 分割交差検定を行った。SVM は線形カーネルで他のパラメータはデフォルト値である。実装には python の機械学習ライブラリの scikit-learn を使用した。

4.3 結果と考察

実験の結果を表 2 に示す。結果として、「研究目的」・「それ以外」の場合、全ての手法で実験目的と比較して、高い値を示している。特に「研究目的」で見ると、手がかり語を用いて分類した場合が高い値を示している。このことから、「研究目的」の場合、目的を表す特徴は表層に出現していると考えられる。しかし「実験目的」の場合、全ての手法で他の目的と比較して、低い値を示しており、特に手がかり語においては、全ての値が最低値である 0 になっている。さらに分類結果として出力したものを観察したところ、そもそも「実験目的」として分類されているパラグラフ自体の数が非常に少なかった。これにより、「実験目的」における特徴は表層には現れず、現状の各手法では「実験目的」の特徴を捉えきれていないと判断できる。

5 おわりに

本研究では論文を対象に、目的を含むパラグラフの推定手法を提案した。また、提案した手法が有効であるかを目的ごとにタグ付けしたデータを使って検証した。結果として、研究目的・それ以外では実験目的に比べ、高い精度を示すことができたが、実験目的に関しては他の目的に比べ、低い精度を示すことになった。今後は各目的の推定精度の向上のために、それぞれの特徴を表すような素性の検討をするとともに、現状では考慮していない文脈を含めた素性の検討や因果関係における目的以外の要素に対する推定に取り組んでいく予定である。

参考文献

[1] 澤村瞳, 小林一郎 “文書内の事象間の関係抽出への取り組み” 人工知能学会 第 27 回, 2013

[2] 青野壮志, 太田学. “要因検索による因果関係ネットワークの構築と因果知識の獲得”. DEIM Forum 2010 B9-1, 2010

[3] 乾孝司, 乾健太郎, 松本裕治. “接続標識「ため」に基つく文章集合からの因果関係知識の自動獲得” 情報処理学会論文誌, Vol. 45, No. 3, pp. 919933, 2004.

[4] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237 (2004.)