

TF-Inverse Component Frequency と GRU による論文構成自動推定法

遠藤 立

February 16, 2020

Contents

1	はじめに	2
2	論文構成の自動推定	3
2.1	χ^2 値	4
3	提案手法	6
3.1	コーパス作成	7
3.2	特徴量	7
3.2.1	TF-ICF	7
3.2.2	本研究で用いている特徴量	10
4	評価実験	13
4.1	実験結果	13
4.2	考察	13
5	むすび	17

1 はじめに

アカデミックライティングでは、文章を適切に構造化することで読者に対して効果的に情報を伝達することが重要である [1,2,3]。アカデミックライティングにおける重要な文章構造として論文構成が知られている [4]。論文構成とは「研究背景」、「関連研究」、「本研究の目的」といった論文の構成要素 (以下、論文要素) の系列を指す。

そのため、これまでに論文構成の構築を支援する多数のシステムが開発されてきた。例えば、O'Rourke and Calvo[5] はパラグラフ間の二元性を可視化するシステムを開発している。Fujimoto and Matsuo[6] は、論文を段落単位に自動分割し、その配列の変更や文章の改定を他の利用者との議論を通じて行うシステムを開発している。更に、Kunichika et al.[7] は、利用者による文章の自由なグループ化と、そのグループの階層的な構造化を支援するシステムを開発している。また、西村ら [8]、甲斐ら [9] は、文章の表層表現に基づいて論文構成を解析し、表示するシステムを開発している。これらのシステムは論文構成の画一的な形式化のみを支援しており、多様な利用者の意図を表現することが難しい。これらの問題に対して、Aluisio and Oliveira[4] や岩田 [10] らは形式的なアプローチを用いてこの問題の解決に取り組んでいる。しかし、形式的なアプローチを用いた場合、無数のテンプレートの作成が必要となるため、このアプローチは現実的ではないと考えられる。この問題を解決するために、宇都・植野 [11] は初心者であっても容易に論文構成を構築できるように確率的なアプローチを用いた論文構成構築支援システムを開発している。しかし、このシステムはユーザー自身が論文構成の入力を行う必要がある。

近年では、論文構成の自動推定技術が提案されている。論文構成の自動推定とは、文や文集合を定義された論文要素に自動的に割り当てることである。Teufel et al.[12,13] はナীবベイズ (Naive Bayes;NB) を用いて論文構成の自動推定が行えることを示している。Guo,Korhonen and Poibeau[14,15] はサポートベクターマシン (Support Vector Machine;SVM)、条件付き確率場 (Conditional Random Field;CRF)、NB を用いて生物医学分野の論文に対して論文構成自動推定精度の評価実験を行なっている。Merity et al.[16] や、Mochales and Moens[17] は、最大エントロピーモデル (Maximum Entropy models;ME) を利用して論文構成の自動推定を行なっている。Widyantoro and Amin[18] は、論文構成の自動推定を行うことで、関連研究の要約をする手法を提案しており、NB や決定木を用いるこ

とで高精度に推定できることを示している。Liu[19] は、特徴量に Word2Vec を用いることで論文構成の推定精度が向上することを示している。これらの論文では特徴量としてユニグラム及びバイグラムを採用しており、頻出上位の単語や出現している全ての単語を選択している。しかし、論文要素ごとに頻出の単語が異なること、論文要素の中でも出現回数が多い論文要素と少ない論文要素で差がある。以上より、ユニグラムやバイグラムを特徴量として利用する際に、頻出上位の単語や出現した全ての単語を選択することは不適切であると考えられる。そこで、Gustavo and Val'eria[20] や Hirohata and Okazaki et al.[21] は、 χ^2 値を用いて単語の重み付けを行っており、論文要素ごとに重みを計算し、各論文要素から重みの大きいユニグラムやバイグラムを特徴量として選択している。しかし、単語によっては複数の論文要素にかけて重みが大きい単語があり、そのような単語は特徴量として適切ではないと考えられる。

そこで、本研究では論文構成の自動推定を行うにあたって、特徴量として利用する単語を選択する手法を提案する。本研究では単語を選択する手法として、自然言語処理において単語の重みづけに利用される TF-IDF(Term Frequency-Inverse Document Frequency) 手法 [22] のアイデアからインスパイアされた手法として Term Frequency-Inverse Component Frequency(TF-ICF) を提案する。提案手法では、ある単語の出現頻度が最大の論文要素 L の出現頻度 TF と、その単語の各論文要素に対する逆頻度 ICF との積 $TF \cdot ICF$ で重み付けを行なっている。提案手法で選択された単語は、論文要素 L の推定に効果的であると考えられる。本研究では、提案手法で選択された特徴量を用いて、日本語論文の序論に対して論文構成の自動推定を行う。論文構成の推定を行うにあたり論文要素として「研究領域の背景」、「関連研究」、「本研究の目的」、「本研究の手法・評価・利点」、「本論の構成」を定めた。また、本研究の目的に利用できる公開コーパスは存在しないため、本研究では独自にコーパスを作成した。推定手法として CRF と時系列データの推定に有効とされるディープラーニングの一種である Long-Short Term Memory(LSTM), Bi-Directional LSTM(Bi-LSTM), Gated Recurrent Unit(GRU), Recurrent Neural Network(RNN)[22,23,24,25] を利用し、論文構成の推定精度の実験を行う。

2 論文構成の自動推定

近年、論文構成の自動推定技術が提案されている。論文構成の自動推定とは、文や文集合を定義された論文要素に自動的に割り当てることである。先行研究で

定義されている論文要素として, Teufel et al., Merity et al., Liu は [12,13,16,19] は,「本研究の目的」,「本研究の説明」,「研究背景」,「先行研究を肯定している文」,「先行研究を否定している文」,「先行研究について単純に述べている文」,「セクションの内容の説明」, の七種類, Guo,Korhonen and Poibeau[14,15] は,「研究背景」,「提案手法」,「結果」,「結論」の四種類, Gustavo and Val'eria[20] は「研究背景」,「本研究の目的」,「提案手法」,「結果」,「結論」,「先行研究」,「今後の展望」の七種類に分類する手法を提案している. また, Hirohata and Okazaki et al.[21] は, 科学的論文は論文要素として主に,「研究背景」,「本研究の目的」,「提案手法」,「結果」,「結論」,「導入」の六種類に分けられることを示している. また各手法において論文要素推定精度の F 値はそれぞれ, Teufel et al.[12,13] は NB を用いることで 50%, Guo,Korhonen and Poibeau[14,15] は SVM を用いることで 90%, Merity et al.[16] は ME を用いることで 97.9%, であることが示されている. これらの論文では特徴量としてユニグラム及びバイグラムを採用しており, 頻出上位の単語や出現している全ての単語や頻出の単語をユニグラム, バイグラムとして採用している. しかし, 論文要素ごとに頻出の単語が異なること, 論文要素の中でも出現回数が多い論文要素と少ない論文要素で差がある. 以上より, ユニグラムやバイグラムを特徴量として利用する際に, 頻出上位の単語や出現した全ての単語を選択することは不適切であると考えられる. そこで, Gustavo and Val'eria[20] や Hirohata and Okazaki et al.[21] は, χ^2 値を用いて単語の重み付けを行っており, 論文要素ごとに重みを計算し, 各論文要素から重みの大きいユニグラムやバイグラムを特徴量として選択している. 各手法において論文要素推定精度の F 値はそれぞれ, Gustavo and Val'eria[20] は LSTM を用いた Encoder-decoder モデルを用いることで 94.68%, Hirohata and Okazaki et al.[21] は CRF を用いることで 93.0%であることが示されている.

2.1 χ^2 値

Gustavo and Val'eria[20] や Hirohata and Okazaki et al.[21] は, χ^2 値を用いて単語の重み付けを行っている. ある文の状態 S を, その文の論文要素が L_j である ($cl = 1$) または, 論文要素が L_j ではない ($cl = 0$) の二種類と, ある文が単語 w_i を含む ($cw = 1$) または, 単語 w_i を含まない ($cw = 0$) の二種類とで分けられる計四種類の状態で表すとする. このときの各状態における文の総数を表すクロス表を Table 1 に示す. Table 1 は, 例えば, 状態 $S(1, 1)$ を表す, 論文要素

Table 1: 各状態の文の総数を表すクロス表

w/L	論文要素 L_j である ($cl = 1$)	論文要素 L_j ではない ($cl = 0$)	合計
w_i を含む ($cw = 1$)	$N(1, 1)$	$N(1, 0)$	$N(1,)$
w_i を含まない ($cw = 0$)	$N(0, 1)$	$N(0, 0)$	$N(0,)$
合計	$N(, 1)$	$N(, 0)$	N

が L_j かつ単語 w_i を含む文の実際の総数は $N(1, 1)$ であり, 論文要素に関わらず単語 w_i を含む文の総数を $N(1, 1) + N(1, 0) = N(1,)$ と表現することを示している. ここで, 論文要素 L_j と単語 w_i が独立であると仮定したときに期待される状態 $S(cw, cl)$ における文の総数を $E(cw, cl)$ と表すとすると,

$$E(cw, cl) = N(cw,) \cdot \frac{N(, cl)}{N} \quad (1)$$

となる. このとき, 実際の状態 $S(cw, cl)$ の文の総数 $N(cw, cl)$ と, 論文要素 L_j と単語 w_i が独立であると仮定したときに期待される状態 $S(cw, cl)$ における文の総数 $E(cw, cl)$ がどの程度乖離しているかで表される指標として χ^2 値が定義されており, 単語 w_i と論文要素 L_j の χ^2 値は

$$\chi^2(i, j) = \sum_{cw=0,1} \sum_{cl=0,1} \frac{(N(cw, cl) - E(cw, cl))^2}{E(cw, cl)} \quad (2)$$

で表される. $N(cw, cl)$ と $E(cw, cl)$ の差が大きければ大きいほど論文要素 L_j と単語 w_i は関係していると考えることができる. 従って, $\chi^2(i, j)$ が大きいならば w_i は論文要素 L_j の分類に適していると考えることができる.

しかし, 論文要素ごとの文の数は大きく異なっており, 全体の文の総数における割合が小さい論文要素として L_p, L_q を仮定すると L_p, L_q で共に単語 w_a が頻出だった際に w_a は L_p 及び L_q の分類に適しているとは限らない. 実際に「本研究の目的」と「提案手法」といった論文要素間では頻出語句が被る傾向があるため, 定義されている論文要素の内容によっては分類精度を低下させてしまう可能性がある.

3 提案手法

本研究では、日本語論文の序論に対して論文構成の自動推定を行う。前述した通り、先行研究では複数の論文要素が定義されている。先行研究では、論文全体の文や文集合に対して論文構成の自動推定を行なっているので、「結果」や「結論」、「セクションの内容の説明」といった論文要素を定義する必要はないと考えられる。そこで、本研究では論文要素として、「研究領域の背景」、「関連研究」、「本研究の目的」、「本研究の手法・評価・利点」、「本論の構成」の五種類を定め、Table 2 に示した。また、本研究の目的に利用できる公開コーパスは存在しないため、本研究では独自にコーパスを作成した。論文要素は論文の分野によって大きく異なるため、異なる分野の論文を混ぜることは望ましくない。本研究では電子情報通信学会論文誌 D「情報システム」関連の日本語の優良論文 100 件に対してラベリングを行った。また、本研究では論文構成の自動推定をするにあたって、特徴量として利用する単語を選択する手法を提案する。本研究では自然言語処理において単語の重みづけに利用される TF-IDF(Term Frequency-Inverse Document Frequency) 手法のアイデアからインスパイアされた手法として、TF-ICF を提案する。推定手法として CRF に加えて時系列データの推定に有効とされるディープラーニングの一種である LSTM, Bi-LSTM, GRU, RNN を利用する。

Table 2: 本研究で使用するラベル

ラベル	内容	詳細
BKG	研究領域の背景	研究領域の社会的意義や課題について述べている文
REL	関連研究	本研究に関連する先行研究のレビューを述べている
OBJ	本研究の目的	本研究の目的を明確に述べている
CON	本研究の手法・評価・利点	本研究の手法, 評価, 利点のいずれかについて述べている
STR	本論の構成	2 章以降で述べる内容及び章番号を述べている

3.1 コーパス作成

本研究で成したコーパスの例として論文 [26] へのラベリングの一部を Table 3 に示す. 本コーパスは Table 3 のように一つの文に対して一つのラベルを付与したデータとして定義した. また, Table 4 に論文要素ごとの文の総数と割合を示す.

本システムはこのコーパスを訓練データとして, 本研究で作成した特徴量を用いて論文構成要素の推定を行う.

3.2 特徴量

本研究で作成したコーパスに対して, 論文要素ごとの χ^2 値の重み上位のユニグラム 5 単語を Table 5, バイグラム 5 単語を Table 6 に示す.

Table 5, Table 6 が示す通りユニグラムやバイグラムが複数の論文要素にまたがって採用されてしまっている単語が多く, これらの単語は論文要素の分類に不適切であると考えられる. そこで, 本研究ではユニグラムやバイグラムの選択手法として TF-ICF を提案する.

3.2.1 TF-ICF

提案手法は, 自然言語処理において文書分類を行う上で, 単語の重みづけに利用される TF-IDF 手法のアイデアからインスパイアされた手法である. まず, TF-IDF 法 [27] を紹介する. 分類すべき文書を $D_l (l = 1, 2, \dots, n)$ とする. 文書 D_l に含まれる単語 w_k の出現回数を $AD(k, l)$, 文書 D_l に含まれる全単語数を TD_l とすると

$$TF(k, l) = \frac{AD(k, l)}{TD_l} \quad (3)$$

は, 文書 D_l における w_k の出現頻度を表している. $TF(k, l)$ の値を大きくする単語 w_k は, 文書 D_l の分類に寄与すると考えられる. しかし, 単語 w_k が多くの文書に出現している場合, w_k は文書分類を行う上で不適切な単語であると考えられる. そこで, 単語 w_k が 1 回以上出現する文書数を ND_k とする. 単語 w_k が登場する文書の逆頻度を

$$IDF_k = \log \frac{n}{ND_k + 1} \quad (4)$$

Table 3: 本研究で作成したコーパス中の論文 [26] のラベリング例

文番号	ラベル	文
1	BKG	機械翻訳においても、また、翻訳者による翻訳においても、高い訳質を保証するためには大規模で正確な対訳辞書が必要不可欠である。
2	BKG	しかし、人手によって対訳辞書を作成し、継続的に収録語数を増やし辞書を維持・管理していく作業は膨大な時間と労力を要する。
3	REL	そこで、自然言語処理分野においては、多様なテキストデータを情報源として、対訳辞書を自動若しくは半自動的に作成する技術に関する研究が行われてきた。
4	REL	例えば、複合語専門用語の構成要素の訳語を連結して訳語の候補を生成し（要素合成法）、ウェブから収集した目的言語の専門分野コーパスを用いて、生成された訳語候補を検証する手法 [3]、検索エンジンから訳語が併記された文書を収集し、訳語対を獲得する手法 [4] 等がある。
		(中略)
9	OBJ	以上の背景のもとで、本論文では、句に基づく統計的機械翻訳モデル [5] により学習されるフレーズテーブル、及び、既存の対訳辞書を用いる要素合成法 [3] という、性質の異なる 2 種類の知識源（大規模対訳文対から統計的に学習された訳語知識、及び、既存の対訳辞書）を用いる手法を併用して、対訳文から専門用語対訳辞書を獲得する手法を提案する。
10	CON	本論文の手法では、まず、専門用語対訳辞書獲得の情報源として用いる日英対訳文対に対して、句に基づく統計的機械翻訳モデルを適用することにより、フレーズテーブルを学習する。
		(中略)
30	STR	以下に本論文の構成を述べる。
31	STR	2. で、本論文で用いた日英対訳特許文について説明し、 3. で、本論文で用いた訳語推定手法について述べる。
		(中略)

Table 4: 論文要素ごとの文の総数と割合

ラベル	文の数	コーパス内の割合 (%)
BKG	501	21
REL	1004	42
OBJ	145	6.1
CON	536	22.4
STR	202	8.5
合計	2388	100

Table 5: χ^2 値重み上位のユニグラム

	BKG	REL	OBJ	CON	STR
1	本	本	本	本	述べる
2	手法	論文	そこで	用いる	章
3	提案	ら	提案	しかし	では
4	企業	ない	論文	示す	説明
5	近年	述べる	研究	候補	最後

Table 6: χ^2 値重み上位のバイグラム

	BKG	REL	OBJ	CON	STR
1	などの	本論文	提案する	提案手法	章で
2	本論文	ている	を提案	ている	について述べる
3	では	本研究	本研究	本手法	について説明
4	ている	提案する	論文で	を用いる	関連研究
5	日本語の	論文で	そこで本	情報量	構成は

で計算する。このとき、 $TF - IDF(k, l) = TF(k, l) \cdot IDF_k$ を文書 D_l に対する単語 w_k の重みとし、この重みが上位の単語 w_k は、文書 D_l の分類に寄与すると考えられる。

続いて、提案手法である TF-ICF について説明する。分類すべき論文要素のラベルを $L_j (j = 1, 2, \dots, m)$ とする。論文要素 L_j に含まれる単語 w_i の出現回数を $AC(i, j)$ 、論文要素 L_j に含まれる全単語数を TC_j とすると

$$tf(i, j) = \frac{AC(i, j)}{TC_j} \quad (5)$$

は、論文要素 L_j における w_i の出現頻度を表している。 $tf(i, j)$ の値を大きくする単語 w_i は、論文要素 L_j の分類に寄与すると考えられる。ここで、 $max = \arg \max_{1 \leq j \leq m} tf(i, j)$ とすると、 $tf(i, max)$ は w_i は特に論文要素 L_{max} において最も出現頻度が多いということを意味しており、 w_i が論文要素 L_{max} の分類に寄与することを期待できる。ここで、 L_{max} 以外の論文要素の単語 w_i の出現頻度に比べて $tf(i, max)$ が大きく離れていることが示せば w_i は出現頻度の点で L_{max} の分類に優れていると考えられる。そこで、単語 w_i の出現頻度における重みを

$$TF_i = \frac{tf(i, max)}{\frac{1}{m} \sum_{j=1}^m tf(i, j)} \quad (6)$$

とする。また、単語 w_i が多くの種類の論文要素に出現している場合、 w_i は論文要素の分類を行う上で、不適切な単語であると考えられる。そこで、閾値 N としたとき、単語 w_i が N 回より多く出現するラベル数を $NC(i, N)$ とする。単語 w_i が N 回以上登場する論文要素の逆頻度を

$$ICF(i, N) = \log \frac{m}{NC(i, N) + 1} \quad (7)$$

で計算する。このとき、 $TF - ICF(i, N) = TF_i \cdot ICF(i, N)$ を単語 w_i の重みとし、この重みが上位の単語を選択することで論文構成の自動推定精度の向上に寄与できると考えられる。

本研究で作成したコーパスに対して、閾値 $N (0 \leq N \leq 9)$ ごとにまとめた TF-ICF 重み上位の単語ユニグラム及びバイグラムを Table 7, Table 8 に示す。Table 5 と Table 7 を比較すると、 χ^2 乗値の重み上位で重複して存在している単語は、確かに TF-ICF 上位の単語に選択されていないことがわかる。

3.2.2 本研究で用いている特徴量

本研究で用いている特徴量を Table 9 に示した。本研究で作成したコーパ

Table 7: TF-ICF 重み上位のユニグラム

	N=0	N=1	N=2	N=3	N=4	N=5	N=6	N=7	N= 8	N=9
1	で	で	で	で	で	で	で	で	で	で
2	ユーザ	行う	行う	行う	行う	行う	処理	処理	処理	処理
3	処理	ある	可能	可能	可能	可能	行う	行う	行う	行う
4	行う	その	ある	ある	ある	ある	可能	可能	ある	可能
5	可能	する	その	その	その	その	ある	ある	文	ある

Table 8: TF-ICF 重み上位のバイグラム

	N=0	N=1	N=2	N=3	N=4	N=5	N=6	N=7	N=8	N=9
1	を行う	ことが	ことが	ことが	ことが	ことが	ことが	ことが	ことが	ことが
2	ことが	を行う	を行う	を行う	を行う	を行う	を行う	ことで	ことで	ための
3	ことで	するため	ている	するため	するため	するため	するため	を行う	するため	ている
4	行うこと	情報を	ことは	ている	ている	ている	ている	するため	ている	ことは
5	するため	ている	これらの	ことは	すること	すること	すること	ている	ことは	すること

Table 9: 本研究で使用する特徴量

特徴量	詳細
ユニグラム	TF-ICF の重み上位 100 語
バイユニグラム	TF-ICF の重み上位 100 語
文番号	文の先頭から数えて何文目か
段落番号	何段落目に含まれるか
箇条書きか否か	—
箇条書きの先頭か否か	—
段落の開始	—
段落の終了	—
単語数	文に含まれている単語数

スに対して TF-ICF を利用してユニグラム及びバイグラムの重みの大きい単語上位 100 語を特徴量として利用した. Teufel et al.[12,13], Guo,Korhonen and Poibeau[14,15], Merity et al.[16] は文章中の位置を特徴量として採用している. 実際に,「研究背景」は文頭にくる場合が多く,「本論の構成」は末尾に書かれるといった文章中の位置から求められる論文要素の情報は非常に大きい. そこで本研究では, 特徴量として「文番号」,「段落番号」,「箇条書きか否か」,「箇条書きの先頭か否か」,「段落の開始」,「段落の終了」を採用した. Teufel et al.[12,13], Guo,Korhonen and Poibeau[14,15], Merity et al.[16], Mochales and Moens[17], Widiantoro and Amin[18], Hirohata and Okazaki et al.[21] は文章中に含まれている単語数を特徴量として採用している. 特に本論の目的について述べる文は簡素で短い文で書かれることが多い. 本研究も同様に特徴量として単語数を採用した. 単語数は OBJ の論文要素であるか否かを推定するのに大きく貢献してくれると期待できる.

本研究ではこれらの特徴量を用いて CRF, 及びディープラーニングの一種である LSTM, Bi-Directional LSTM, GRU, RNN を利用して論文構成の推定精度の実験を行う.

4 評価実験

本研究では, Table 9 に示した特徴量を用いて CRF, 及びディープラーニングの一種である LSTM, Bi-Directional LSTM, GRU, RNN を利用して論文構成の推定精度の実験を行う. TF-ICF の閾値 $N(0 \leq N \leq 9)$ を変更し, 論文構成の推定実験を行うことで本コーパスに対して適切な特徴量を求める. またユニグラム, バイグラムを本研究で作成したコーパスの頻出 100 語に変更した場合及び χ^2 値によるユニグラム, バイグラムを用いた単語に変更した場合で評価実験を行なった. 本研究で作成したコーパスを用いて, 10 分割交差検証により実験を行なった.

4.1 実験結果

推定精度の F 値を結果としてそれぞれ Table 10~Table 14 に示す. Table 10~Table 14 より LSTM, GRU, RNN については, TF-ICF の閾値の値として $N = 3$ が全てのラベル及び全体の推定精度において最も推定精度が高いことが示せた. また, CRF についても OBJ, CON, 及び全体の推定精度において $N = 3$ が最も精度が高いことが示せた. Bi-Directional LSTM, については STR を除く全てのラベル及び全体の推定精度において最も推定精度が高いことが示せた. また, 手法同士を $N = 3$ において比較すると STR を除いた全てのラベル及び全体の推定精度において GRU が最も推定精度が高いことが示せた. STR については CRF が最も精度が高いことが示せた. 以上より, 本コーパスにおいて TF-ICF を閾値 $N = 3$ で用いて, GRU を利用した場合が最も推定精度が高くなることが示せ, 90%を超える推定精度が示せた.

4.2 考察

LSTM のモデル図を Figure1, GRU のモデル図を Figure2 に示す. LSTM は RNN に比べて長期の系列優れるモデルである. LSTM の時点 t における時系列状態 x_t , 時点 $t-1$ におけるセルの状態 c_{t-1} と隠れ層の状態 h_{t-1} を入力として, 各ゲート及び出力は以下の式で求められる.

$$f_t = \sigma(W_f x_t + U_f h_{t-1}) \quad (8)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1}) \quad (9)$$

Table 10: CRF

N/ラベル	BKG	REL	OBJ	CON	STR	ALL
0	0.6966	0.7252	0.5213	0.6290	0.8876	0.6920
1	0.7305	0.7471	0.5454	0.6612	0.8672	0.7103
2	0.7036	0.7341	0.5516	0.6531	0.8695	0.7024
3	0.7129	0.7526	0.5745	0.6874	0.8982	0.7251
4	0.6925	0.7379	0.5158	0.6382	0.8736	0.6916
5	0.6879	0.7453	0.5253	0.6592	0.8682	0.6972
6	0.7138	0.7391	0.5399	0.6816	0.8985	0.7146
7	0.6765	0.7309	0.5647	0.6435	0.9110	0.7053
8	0.6739	0.7182	0.5529	0.6485	0.8834	0.7036
9	0.6759	0.7257	0.5469	0.6570	0.8836	0.7118
頻出	0.6575	0.7426	0.4852	0.6520	0.7653	0.6605
χ^2 値	0.6832	0.7616	0.5304	0.6753	0.7891	0.6879

Table 11: LSTM

N/ラベル	BKG	REL	OBJ	CON	STR	ALL
0	0.6642	0.6367	0.4588	0.4903	0.7776	0.6205
1	0.6576	0.6451	0.4881	0.5162	0.7797	0.6229
2	0.6750	0.6866	0.4616	0.5402	0.7654	0.6484
3	0.8739	0.8750	0.6073	0.7768	0.7942	0.8336
4	0.6577	0.6784	0.5105	0.5786	0.7806	0.6557
5	0.6404	0.6569	0.5215	0.5317	0.7801	0.6314
6	0.6411	0.6492	0.4598	0.5544	0.7773	0.6309
7	0.6652	0.6537	0.4655	0.5375	0.7262	0.6310
8	0.6515	0.6242	0.4613	0.5210	0.7317	0.6110
9	0.6444	0.6283	0.4996	0.5487	0.7583	0.6204
頻出	0.6706	0.6501	0.4767	0.5064	0.6556	0.6199
χ^2 値	0.649	0.6616	0.4239	0.5227	0.6116	0.6073

Table 12: Bi-LSTM

N/ラベル	BKG	REL	OBJ	CON	STR	ALL
0	0.6765	0.6658	0.4893	0.5396	0.8470	0.6511
1	0.6656	0.6797	0.5190	0.5617	0.8453	0.6589
2	0.6868	0.6959	0.4568	0.5994	0.8471	0.6751
3	0.8453	0.8824	0.6429	0.8633	0.8590	0.8568
4	0.6665	0.6909	0.4908	0.5657	0.8391	0.6640
5	0.6665	0.6662	0.5223	0.5446	0.8754	0.6527
6	0.6705	0.6875	0.4606	0.5250	0.8194	0.6512
7	0.6427	0.6693	0.4342	0.5728	0.8418	0.6457
8	0.6805	0.6890	0.5099	0.5812	0.8226	0.6715
9	0.6461	0.6585	0.5066	0.5661	0.8427	0.6475
頻出	0.6511	0.6814	0.4682	0.5483	0.7126	0.6428
χ^2 値	0.643	0.6774	0.4321	0.5444	0.7391	0.643

Table 13: GRU

N/ラベル	BKG	REL	OBJ	CON	STR	ALL
0	0.6669	0.6628	0.5165	0.5549	0.8210	0.6520
1	0.6797	0.6437	0.4704	0.5386	0.8649	0.6370
2	0.6928	0.6958	0.5492	0.5975	0.7953	0.6780
3	0.9156	0.9418	0.7395	0.8729	0.8181	0.9018
4	0.6648	0.6845	0.5884	0.5647	0.7869	0.6649
5	0.6378	0.6358	0.5104	0.5730	0.8043	0.6335
6	0.652	0.6622	0.5158	0.5224	0.8017	0.6408
7	0.6468	0.6341	0.5121	0.5744	0.7597	0.6271
8	0.6604	0.6609	0.5297	0.5194	0.7796	0.6373
9	0.6678	0.6594	0.5386	0.5375	0.7970	0.6465
頻出	0.6744	0.6490	0.5101	0.5402	0.7428	0.6364
χ^2 値	0.6262	0.6556	0.5287	0.5316	0.6952	0.6202

Table 14: RNN

N/ラベル	BKG	REL	OBJ	CON	STR	ALL
0	0.6887	0.6428	0.4582	0.521	0.7194	0.624
1	0.6792	0.6275	0.4379	0.4709	0.7236	0.6041
2	0.6893	0.6882	0.501	0.5765	0.7329	0.6594
3	0.8303	0.8177	0.5468	0.7472	0.7836	0.7857
4	0.6974	0.6700	0.5148	0.5228	0.7306	0.6477
5	0.6778	0.6552	0.4777	0.5115	0.7182	0.6258
6	0.6604	0.6540	0.4486	0.5508	0.7631	0.6383
7	0.6580	0.6385	0.4366	0.5665	0.7238	0.6255
8	0.6648	0.6361	0.4242	0.5547	0.7381	0.6283
9	0.6449	0.6519	0.5014	0.5796	0.7377	0.6332
頻出	0.7050	0.6581	0.4741	0.5035	0.6551	0.6269
χ^2 値	0.6511	0.6586	0.3961	0.5453	0.6144	0.6149

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1}) \quad (10)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1}) \quad (11)$$

$$c_t = f_t c_{t-1} + i_t \tilde{c}_t \quad (12)$$

$$h_t = o_t \tanh(c_t) \quad (13)$$

それに対して GRU の時点 t における時系列状態 x_t , 時点 $t-1$ における隠れ層の状態 h_{t-1} を入力として, 各ゲート及び出力は以下の式で求められる.

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \quad (14)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \quad (15)$$

$$\tilde{h}_t = \tanh(W x_t + r \odot U h_{t-1}) \quad (16)$$

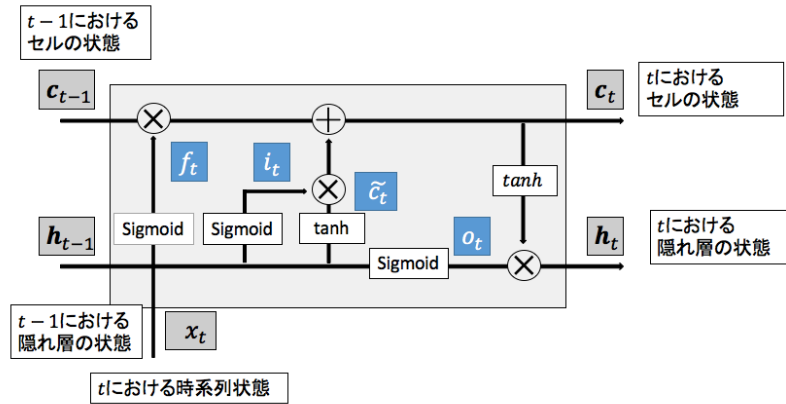


Figure 1: LSTM のモデル図

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_{t-1} \quad (17)$$

式に含まれる U や W は、学習すべきパラメータである。式にある通り GRU は LSTM に比べてパラメータが少ないため、学習用のデータセットが少ない場合に LSTM に比べて推定精度が高くなると考えられる。従って、本研究のような十分なデータセットを用意するのが困難な場合に GRU は優れた推定精度を示すことができると考えられる。

5 むすび

本研究では日本語論文の序論に対して論文構成の自動推定を行なった。論文構成の推定を行うにあたり論文要素として「研究領域の背景」、「関連研究」、「本研究の目的」、「本研究の手法・評価・利点」、「本論の構成」を定めた。また、本研究の目的に利用できる公開コーパスは存在しないため、本研究では電子情報通信学会論文誌 D「情報システム」関連の日本語の優良論文 100 件に対してラベリングを行い、独自にコーパスを作成した。また、本研究では論文構成の自動推定をするにあたって特徴量として利用する単語ユニグラム、バイグラムを選択する手法として TF-ICF を提案した。また、推定手法として CRF, LSTM, Bi-Directional LSTM, GRU, RNN を用いて F 値の推定精度実験を行なった。

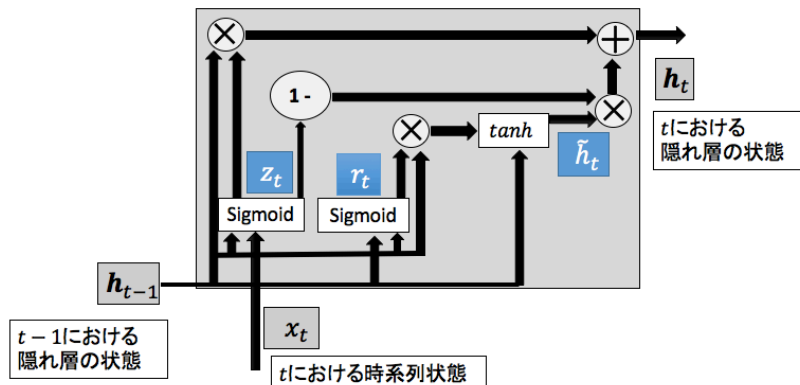


Figure 2: GRU のモデル図

結果としては、TF-ICF を閾値 $N = 3$ を利用した場合、全ての手法において全体の推定精度が最も高いことが示せ、CRF, Bi-LSTM を除く全ての手法で全てのラベルの推定精度が最も高いことが示せた。Bi-LSTM は STR を除く全てのラベル、CRF においては OBJ, CON のラベルで閾値 $N = 3$ が最も推定精度が高いことを示せた。閾値 $N = 3$ で各手法を比較した場合、STR を除く全てのラベル及び全体の推定精度において GRU が最も推定精度が高いことが示せた。STR については CRF が最も精度が高いことが示せた。以上より、TF-ICF を閾値 $N = 3$ で用いて、GRU を利用した場合が最も推定精度が高くなることが示せ、90%を超える推定精度が示せた。

今後は、ユーザーが入力した論文の各文に対して提案を用いて論文構成の自動推定を行い、出力された論文構成が不適切な論文構成であった際にフィードバックを与えるシステムの構築を考えている。提案を用いることで、高精度に論文構成を自動推定を行うことができるため、ユーザー自身が論文構成を入力する必要がない論文構成支援システムを開発できると考えられる。

References

- [1] N.Kando, "Text-level structure of research papers: Implications for text based information processing systems," 19th Annual BCS-IRSG Col-

lowium on IR,Aberdeen,UK,April 1997

- [2] C.M. Neuwirth,D.S. Kaufer, R. Chimera, and T.Gillespie,"The notes program: A hypertext application for writing from source texts,"Proc. Hypertext'87,pp121-142,1987
- [3] H. Shibata and K.Hori,"A framework to support writing as design using multiple representations,"Proc. 5th Asia Pacific Conference on Computer Human Interaction,2002
- [4] S.M. Alusio and O.N. Oliveira, Jr., "A detailed schematic structure of research paper introduction: samiento del Lenguaje Natural,1996."
- [5] S.T. O'Rourke and R.A. Calvo, "Analyzing semantic flow in academic writing," Proc. 2009 Conference on Artificial Intelligence in Education, 2009.
- [6] T. Fujimoto and T. Matsuo, "An approach to cooperative research paper writing/edit method," IJCSNS International Journal of Computer Science and Network Security, vol.7, no.3, pp.15-20, March 2007.
- [7] H. Kunichika, C. Miyazaki, Y. Matsuki, R. Chartrand and A. Takeuchi, "An intelligent partner for organizing a paragraph," Proc. 2009 Conference on Artificial Intelligence in Education,2009
- [8] 西村健士, 島津秀雄, "特定表現の重点解析による科学技術論文構造化手法," 情報学研報, 情報基礎研究報告会 " vol.93.no39,pp59-62,1993.
- [9] 甲斐郷子, 中村順一, 吉田 將, "表層表現に基づく文章構造解析を利用した論文改定支援システムの試作と評価," 情処学研報, 自然言語処理研究会報告, vol.95,no.27,pp.79-84,1995.
- [10] 岩田芳明, 山村 毅, 大西 昇, "マークアップ方式による文章作成システム," 信学技報, NLC97-60,1998.
- [11] 宇都雅輝・植野真臣 (2011) ベイズ符号を用いた論文構成支援システム. 電子情報通信学会論文誌 D , Vol. J94-D, No. 12, pp.2069-2081.
- [12] Simone Teufel "Argumentative Zoning for Improved Citation" Computing Attitude and Affect in Text: Theory and Applications pp.159-169

- [13] Simone Teufel and Marc Moens "Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status Indexing" *Computational Linguistics*, Volume 28, Number 4, December 2002, pp.409-445
- [14] Yufan Guo et. al. (2011) Comparison and user-based evaluation of models of textual information structure in the context of cancer risk assessment. *BMC Bioinformatics*, Vol.12, No. 69, pp.1-18
- [15] Yufan Guo et al.(2011) A Weakly-supervised Approach to Argumentative Zoning of Scientific Documents. *Conference on Empirical Methods in Natural Language Processing*, pp. 273 - 283.
- [16] Stephen Merity and Tara Murphy and James R. Curran "Accurate Argumentative Zoning with Maximum Entropy models" *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, ACL-IJCNLP 2009*, pages 19–26
- [17] Raquel Mochales & Marie-Francine Moens "Argumentation mining" *Springer Science+Business Media B.V. 2011* pp.1-22
- [18] Widyantoro and AminCitation "Sentence Identification and Classification for Related Works Summarization" D. H. Widyantoro, Dan I. Amin, "Citation Sentence Identification and Classification for Related Work Summarization", *Int. Conf. Adv. Comput. Sci. Inf Syst. (ICACISIS)*, 2014. pp.291-296
- [19] Haixia Liu "Automatic Argumentative-Zoning Using Word2vec" *Cornell University Library*, <https://arxiv.org/abs/1703.10152>, (29 March 2017).
- [20] Gustavo Bennemann de Moura and Val'eria Delisandra Feltrim "Using LSTM encoder-decoder for rhetorical structure prediction" *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)* pp.278-283
- [21] Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou and Mitsuru Ishizuka "Identifying Sections in Scientific Abstracts using Conditional Random Fields" pp.381-388

- [22] Mishra et al. "VMAnalyzer: Malware Semantic Analysis using Integrated CNN and Bi-Directional LSTM for Detecting VM-level Attacks in Cloud" 2019 Twelfth International Conference on Contemporary Computing pp.1-6
- [23] W. Bao, J. Yue, and Y. Rao, A deep learning framework for financial time series using stacked autoencoders and long-short term memory, PloS One 12, e0180944 (2017).
- [24] Gamboa J., 2017. Deep Learning for Time-Series Analysis. University of Kaiserslautern, Kaiserslautern, Germany.
- [25] Chung, C. Gulcehre, K. Cho, Y. Bengio, "Gated feedback recurrent neural networks", ICML-15, 2015.
- [26] 渋木英潔 荒木健治 桃内佳雄 栃内香次. 2006. 単語概念の深層格選好に基づく深層格推測手法. 電子情報通信学会論文誌 D. Vol. J89-D No.6 pp1413-1428
- [27] T. Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In International Conference on Machine Learning (ICML), 1997.