ルールベースと機械学習を利用した論文要約からの重要情報抽出

菊井　真[†]　　村田　真樹[‡]　　馬　青[†‡]

[†]龍谷大学大学院理工学研究科

[‡]情報通信研究機構けいはんな情報通信融合研究センター

１．はじめに

これまで書かれてきた論文の中には研究者にとって有用であるにも関わらず、時間的制約によって読まれることのない論文が存在する。そこで我々は、言語処理の論文要約から、重要情報を抽出する研究を行っている。ここでいう重要情報とは表１に示した９種類の分類のことで、例えば、精度表現、応用先、わかったこと、提案手法などである。この重要情報を抽出することにより、論文を読まずして重要な情報を得ることができると考える。先行研究[1]においては、データ量が少ないこともあり精度の向上が困難であった。本研究ではそれに比べ、学習データを増強し若干の性能向上を実現した。また、ルールベースを用いた抽出も試み、さらなる性能向上を実現した。本稿ではこれらの実験結果と問題点等について述べる。

表１　重要表現の種類

重要情報表現		例
表記	分類	
ACCURACY	精度表現	９７％
APPLICATION	応用先	音声認識
CLEAR	わかったこと	研究を進めるなか…がわかった。
FIELD	自然言語処理分野内で主要な分野	機械翻訳、語彙構文解析、抽出
LANGUAGE	言語名	日本語、英語
METHOD	提案手法	トライ構造
NAME	組織・人名	ICOT、サリー大学
NUMBER	数量的成果	５．４万種類の語…知識データを得た。
TECHNIQUE	解くべき問題	知識情報

２．　ベースライン

先行研究では論文要約データを 200 編使用して実験を行っていた。本論文では、この論文要約データを 619 編まで増やし、実験を行った。まずそのベースラインの結果を表２に示す。この実験は抽出すべき重要情報を人手でタグ

表２　論文要約データ 619 編を用いたベースラインでの実験結果

		完全一致			不完全一致			不完全一致（正解数変動）		
重要情報表現	出現総数	再現率	適合率	Ｆ値	再現率	適合率	Ｆ値	再現率	適合率	Ｆ値
精度表現	105	61.90	65.66	63.73	79.05	83.84	81.37	74.30	78.80	76.49
応用先	23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
わかったこと	39	0.00	0.00	0.00	2.56	100	5.00	1.36	53.10	2.65
主要な分野	3689	78.23	79.37	78.80	88.24	89.52	88.87	84.85	86.09	85.46
言語名	687	93.45	96.98	95.18	93.74	97.28	95.48	93.66	97.20	95.40
提案手法	908	12.11	26.19	16.57	14.98	32.38	20.48	13.95	30.15	19.07
組織・人名	106	35.85	86.36	50.67	36.79	88.64	52.00	35.95	86.61	50.81
数量的成果	29	3.45	9.09	5.00	13.79	36.36	20.00	10.02	26.41	14.53
解くべき問題	978	24.85	36.87	29.69	36.50	54.17	43.62	32.70	48.52	39.07
マクロ平均		34.43	44.50	37.74	40.63	64.69	45.20	38.53	56.32	42.61
マイクロ平均		60.17	72.00	65.87	68.85	81.64	74.70	66.11	78.40	71.13

付けした学習データで、YamCha を利用して実験を行った。評価は５分割 Cross-Validation を利用し、素性として先行研究[2]で用いられていたものを付与した。

先行研究では解答データの評価方法として、抽出結果の解答データと正解データとが完全に一致していれば正解、一致していなければ不正解としてきた。しかし抽出対象が句や文などの長い表現の場合、完全に正解と一致していなくても正解と判断してもよいような結果も多く見られた。そこで、本稿では次の新たな２種類の算出方法でも精度を求める。

不完全一致

解答データと正解データとが少しでも一致している部分があれば正解とする。

不完全一致（正解数変動）

解答データと正解データとの一致している文字数により個々の正解数（num ：$0 \le num \le 1$）を変動させる。個々の正解数は

$$num = \frac{2 \times agreementLength}{correctLength + ansLength}$$

で算出する。$agreementLength$ は正解データと解答データが一致している文字数を、$correctLength$ は正解データの文字数を、$ansLength$ は解答データの文字数をそれぞれ表す。

これ以降、これまでの完全一致の場合での算出方法と合わせ、これら３種類で精度を算出する。

３．　重要情報表現

３．１　『精度表現』

先行研究において正規表現パターンを利用した場合の論文要約データ 200 編に対する抽出を試みた。その結果高い精度を得ることができた。ここではまず論文要約データ 619 編での正規表現パターンを利用した場合の抽出精度を表３に示す。

表３　正規表現パターンでの結果

	再現率	適合率	Ｆ値
完全一致	84.76	60.54	70.63
不完全一致	96.19	68.71	80.16
不完全一致（正解数変動）	93.87	67.05	78.22

その結果、高い精度を得ることができたが、この正規表現パターンを YamCha に適用すれば、さらに高い精度を得ることができるのではないかと考えた。そこで、作成された正規表現パターンと一致するパターンが論文要約データに存在した場合、その部分に『精度表現』のパターンであることを意味する素性を学習時に追加した。この実験結果を表４に示す。

表４　正規表現パターン＋Yamcha での結果

	再現率	適合率	Ｆ値
完全一致	69.52	66.36	67.91
不完全一致	82.86	79.09	80.93
不完全一致（正解数変動）	79.06	75.46	77.22

実験の結果、多少ながらベースラインよりは精度を上げることができたが、正規表現パターンより良い精度を出すことができなかった。しかし YamCha の出力結果を見ると、大きく間違った答えは出していない。現在の学習数を考えれば、もし学習数を増やすことができれば精度向上の可能性はあると考える。

３．２『応用先』『わかったこと』『数量的成果』

この３つの分類については極端に低い精度でしか抽出できていない。そこで今回ルールベースでの抽出を試みた。

実験方法は本来なら５CV での実験を行うべきであるが、元々学習数が少ないので評価用データにその３種類の重要情報表現が出現しない可能性があるうえ、もし出現したとしても極端な精度になり公平な計算ができない可能性がある。よって今回は、

『わかったこと』についてのルール
文末表現として「確認された。」「分かった。」「わかった。」「得ることができた。」のいずれかで終了している。
『数量的成果』についてのルール
文中に「…た結果、」「本論文では」「本稿では」のいずれかが存在しており、かつ文中に数字が存在する。
『応用先』についてのルール
文中に「…期待できる。」「…期待される。」「…応用できる。」「…として利用できる。」「…への応用を…」「…に対しても有効である。」「…で用いられている。」が存在する。

図１　close での最高精度のルール

表５　close での実験結果

	完全一致			不完全一致			不完全一致（正解数変動）		
重要情報表現	再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値
わかったこと	31.82	18.92	23.73	86.36	51.35	64.41	75.28	44.76	56.14
数量的成果	22.22	22.22	22.22	77.78	77.78	77.78	55.09	55.09	55.09
応用先	0.00	0.00	0.00	68.42	100.00	81.25	26.65	38.95	31.65

表６　open での実験結果

	完全一致			不完全一致			不完全一致（正解数変動）		
重要情報表現	再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値
わかったこと	17.65	4.76	7.50	64.71	17.46	27.50	56.30	15.19	23.93
数量的成果	0.00	0.00	0.00	10.00	66.67	17.39	7.71	51.43	13.42
応用先	0.00	0.00	0.00	75.00	25.00	37.50	43.73	14.58	21.86

表７　データ①を学習用データに追加しての実験結果

	完全一致			不完全一致			不完全一致（正解数変動）		
重要情報表現	再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値
精度表現	74.29	72.22	73.24	84.76	82.41	83.57	81.84	79.56	80.68
応用先	13.04	100.00	23.08	13.04	100.00	23.08	13.04	100.00	23.08
わかったこと	0.00	0.00	0.00	2.56	9.09	4.00	2.38	8.45	3.72
組織・人名	59.43	82.89	69.23	67.92	94.74	79.12	65.27	91.04	76.03
数量的成果	13.79	16.00	14.81	31.03	36.00	33.33	27.81	32.26	29.87
マクロ平均	32.11	54.22	36.07	39.86	64.45	44.62	38.07	62.26	42.68
マイクロ平均	49.01	66.37	56.38	57.62	78.03	66.29	55.34	74.94	63.66

表８　データ①②を学習用データに追加しての実験結果

	完全一致			不完全一致			不完全一致（正解数変動）		
重要情報表現	再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値
精度表現	73.33	73.33	73.33	82.86	82.86	82.86	80.25	80.25	80.25
応用先	データ出力無し			データ出力無し			データ出力無し		
わかったこと	2.56	10.00	4.08	7.69	30.00	12.24	7.00	27.29	11.14
組織・人名	62.26	83.54	71.35	69.81	93.67	80.00	67.63	90.75	77.51
数量的成果	13.79	28.57	18.60	31.03	64.29	41.86	28.08	58.16	37.87
マクロ平均	30.39	39.09	33.47	38.28	54.16	43.39	36.59	51.29	41.35
マイクロ平均	53.05	71.15	60.78	62.01	83.17	71.05	59.79	80.20	68.51

・論文要約データ前半 300 編でルール作成、前半 300 編に対してルール適用（close）
・close で作成したルールを、後半 319 編に対して適用（open）

の方法で実験を行うこととした。close での実験はまず最高の精度を出せるルールを作成するために行う。そしてそのルールを open で実験することにより評価を行う。最高精度を出したルールを図１に示す。ただし句読点は筆者に

表９　最終結果

	完全一致			不完全一致			不完全一致（正解数変動）		
重要情報表現	再現率	適合率	F値	再現率	適合率	F値	再現率	適合率	F値
精度表現	73.33	73.33	73.33	82.86	82.86	82.86	80.25	80.25	80.25
応用先	13.04	100.00	23.08	13.04	100.00	23.08	13.04	100.00	23.08
わかったこと	2.56	10.00	4.08	7.69	30.00	12.24	7.00	27.29	11.14
主要な分野	78.23	79.37	78.80	88.24	89.52	88.87	84.85	86.09	85.46
言語名	93.45	96.98	95.18	93.74	97.28	95.48	93.66	97.20	95.40
提案手法	16.85	31.74	22.01	20.48	38.59	26.76	18.81	35.43	24.57
組織・人名	62.26	83.54	71.35	69.81	93.67	80.00	67.63	90.75	77.51
数量的成果	13.79	28.57	18.60	31.03	64.29	41.86	28.08	58.16	37.87
解くべき問題	28.22	39.77	33.01	41.31	58.21	48.33	36.84	51.92	43.10
マクロ平均	42.41	60.37	46.60	49.80	72.71	55.50	47.80	69.68	53.15
マイクロ平均	62.58	72.26	67.07	71.07	82.06	76.17	68.17	78.71	73.06

よって用い方が様々であるので全角・半角どちらでも良いとする。

このルールを用いた close での実験結果を表５に、open での実験結果を表６に示す。その結果、「わかったこと」「応用先」について若干精度を向上できた。

４．　データ追加実験

精度が低い理由として、先行研究では論文要約データ中の重要情報表現の出現総数が少ないことを挙げていた。今回論文要約データを 619 編まで増加させたが、まだ出現総数が少ない重要情報表現が存在した。その結果、ベースラインでの実験結果はやはり低い精度しか得ることができなかった。そこで、出現総数の少ない『精度表現』『数量的成果』『組織・人名』『応用先』『わかったこと』の学習データにのみ論文要約データをさらに追加することにより、一時的に出現総数を増加させて実験を行った。追加させたデータは次の２種類である。１つ目は情報処理学会の自然言語処理研究会のデータを電子化して作成した論文要約 1239 編（7514 文）（①）、２つ目は電子情報通信学会の言語理解とコミュニケーション研究会の論文要約データ 424 編（2443 文）（②）である。

データ①を追加した実験結果を表７に、データ①②を追加した実験結果を表８に示す。実験の結果、どの分類についてもベースラインよりも良い精度が得られた。これにより出現総数を増やすと精度を向上させることが可能であることを示すことができた。

５．　実験結果

これまでの改良実験で最も良い精度が得られた実験結果とベースラインでの結果を合わせた最終結果を表９に示す。本研究では学習データの数を増やすことで精度向上を実現できた。さらに正規表現やルールを利用する方法でもさらに精度向上を実現することができた。

６．　おわりに

今回我々は論文要約からの重要情報表現抽出を試みた。実験の結果、重要情報表現の抽出は可能であると考える。しかし、４節でも述べた様にデータ量が依然不足しており、さらなる精度向上の為にはデータの更なる増強が望まれる。さらに今後の課題として、より一層の精度向上を図るために、論文要約に対し形態素解析と構文解析だけでなく意味解析も行い、そこから得られた意味情報を素性として YamCha に適用する試みを考えたい。

参考文献

[1] 菊井真、村田真樹、馬青：言語処理の論文要約からの重要情報抽出、言語処理学会第１１回年次大会、2005

[2] 山田寛康、工藤拓、松本裕治：Support Vector Machine を用いた日本語固有表現抽出、情報処理学会論文誌、Vol.43、No.1、pp.44-53、2002