言語処理の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
[‡]情報通信研究機構

1　はじめに

現在までに様々な研究がなされ、それに応じて多くの論文が書かれているが、個人がこれら全ての論文を読むことは物理的に不可能であると思われる。しかし、これらの中には研究者にとって有意義な論文がいくつも存在する可能性がある。そこで本稿では、論文集から重要と思われる単語や文、及び文章等の情報を自動で抽出する手法を提案する。本手法を適用することにより、大規模な論文集から研究者にとって必要な情報を自動で得ることが出来ると考えられる。また、本研究は、先行論文の要約文集から重要な情報を抽出することにより、これまでどういった研究がなされてきたかを整理するのにも役に立つ。以降、これらの抽出する重要な情報のことを「重要情報」と呼ぶ。関連研究として、文献[1]のように生物学の論文テキストから生物学で重要な表現である蛋白質表現を抽出する研究がある。また、文献[2]では、物理学の論文テキストから物理学にとって有益なデータベースを作ることを目指して、論文テキストから物理学の重要な表現の抽出が試みられている。

先行研究[3][4]では新聞記事を対象にした固有表現抽出で高い精度を出すことに成功している。本稿ではその固有表現抽出の技術を利用して、重要情報抽出を試みる。

2　重要情報抽出

2．1　重要情報表現の定義

先行研究[3]では IREX 日本語固有表現抽出タスクで定義されていた固有表現を使用していた。本研究では論文要約からの重要情報の抽出を行うため、表１に示した９種類の重要情報表現を定義し、使用した。

先行研究[4]では『各固有表現は重なり合うことも入れ子になることもない』としているが、本研究では異なる種類の重要表現については重なり合っても良いことにした。これは以下の例１の「文法解析の技術」のように「自然言語処理で主要な分野」であり、かつ「提案手法」であるといった、複数のものに分類される重要情報表現が存在しうるからである。

表１　重要情報表現の種類

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	分かったこと	１文当り１１．０語で…が判明した。
FIELD	自然言語処理分野内で主要な分野	機械翻訳
LANGUAGE	言語名	日本語
METHOD	提案手法	トライ構造
NAME	組織・人名	ICOT
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

＜例１＞　具体例
中国語入力における＜METHOD＞＜FIELD＞文法解析＜／FIELD＞技術＜／METHOD＞の応用

2．2　機械学習法

本研究では奈良先端科学技術大学院大学で作成された Yamcha[5] を機械学習に使用し、重要表現のまとめあげ状態を表す手法として SVM（サポートベクトルマシン）での固有表現抽出において最も精度が高いとされるチャンクタグ IOB2[4] を用いた。解析方向は文末から文頭とし、SVM では二次の多項式カーネルを利用し、C は１に設定し、one vs rest を利用した。

2．3　実験に使用したデータ

NTCIR１情報検索コンテストで使用されていた論文要約データ[6]の内、研究分野を明確にするため自然言語処理に関する論文要約文章のみを用いた。本稿ではこのような論文要約文を２００編収集した。

これらの論文要約文に対して、以下の例２のように 2.1 節で挙げた表１の分類の重要表現に人手でタグを付けることによって実験に使用する論文要約データを作成した。

＜例２＞　タグ付けされた論文要約データの例
…な文だけを＜FIELD＞生成＜／FIELD＞する＜**LANGUAGE**＞日本語＜／**LANGUAGE**＞＜**TECHNIQUE**＞＜FIELD＞文生成＜／FIELD＞システム＜／**TECHNIQUE**＞について…

３　実験

本研究では複数の種類の重要表現が重なり合うことを認めている。しかし Yamcha では、重要表現が重なり合った状態では実験を行うことができない。そこで本研究では、重要情報表現を種類ごとに分けて実験を行うことにした。

各重要情報表現の抽出精度として再現率と適合率、F-measure を算出し、そこから全重要表現のマクロ平均とマイクロ平均[7]を算出した。なお、本研究では基本的に５分割の Cross-Validation で実験を行った。

３．１　ベースライン

本研究では、先行研究[3]において用いられていた以下の６種類の素性を用いた実験をベースラインとし、実際に実験を行った。

・単語情報　　・品詞情報　　・文字種
・文節内素性　・隣接文節素性　・主辞素性

また、他のベースライン手法として、学習データに出現した重要情報の文字列と同じ文字列をテストデータから抽出する方法も実験に利用した。また、学習データ量を増減させた場合の実験も行った。

ベースラインの実験結果を表２に示す。ここでいう出現総数とは論文要約データ中に存在する各重要情報表現の個数である。

学習データに出現した重要情報の文字列をテストデータから抽出する方法の実験として、次の２種類を行った。一つは、学習データに存在する重要情報の文字列全てを重複を許して抽出する実験（ベースライン A）であり、もう１つは、重なって出現する重要情報の文字列のうち最長のものしか抽出しない実験（ベースライン B）である。例えば、学習データに「意味的曖昧性解消」と「曖昧性解消」という重要表現の文字列が出現したとする。この時、ベースライン A では「意味的曖昧性解消」と「曖昧性解消」の両方をテストデータから抽出するが、ベースライン B では「意味的曖昧性解消」は抽出するが「曖昧性解消」は抽出しない。これらの手法を用いた次の４パターンの実験を行った。

①　ベースライン A
②　ベースライン B
③　ベースライン A＋ベースライン
④　ベースライン B＋ベースライン

上記の③、④では、ベースライン A（もしくは B）かベースラインのどちらかで抽出できたもの全てを結果とした。以上の結果を表３に示す。

学習データ量を増減させた場合の実験では、ベースラインの実験で用いた学習データ量を１／２、１／４にそれぞれ減らして実験を行った。その結果を表４に示す。

表２：ベースラインでの実験結果

	出現総数	再現率	適合率	F-値
精度表現	9	11.11	100	20
応用先	18	0	0	0
分かったこと	18	5.56	11.11	7.41
自然言語処理分野内で主要な分野	1106	77.31	78.44	**77.87**
言語名	243	88.07	91.45	**89.73**
提案手法	205	6.83	24.14	10.65
組織･人名	53	16.98	60	26.47
数量的成果	8	0	0	0
解くべき問題	283	21.20	36.36	26.79
マクロ平均		25.23	44.61	28.77
マイクロ平均		59.39	73.22	65.59

表２より「自然言語処理分野内で主要な分野」「言語名」については高い精度が得られた。その理由として重要情報表現の出現総数が多く、さらに特定の文字列の出現頻度の多い重要情報表現が存在するということが考えられる。これは重要情報表現の総数と出現頻度が多いデータの方が機械学習において効率良く学習できるからである。逆に出現総数と出現頻度が極端に少ない「応用先」「数量的成果」といった分類については極端に低い精度しか得られなかった。

表３よりベースライン A･B よりもベースラインの性能の方が高いことが確認できた。表４よりデータ量の増加に比例して精度も僅かながら向上することが分かった。

３．２　重要表現ごとの実験

精度の低かった分類について、それらの精度を向上させるための追加実験を行った。

３．２．１　「提案手法」「解くべき問題」

「提案手法」「解くべき問題」は論文要約データのタイトル部分に出現することが多い。そこで『タイトル部分である』ことを意味する素性『TITLE』を追加することにより、精度向上を目指した。また、素性『TITLE』を付与し

表３：ベースラインとの比較

	ベースライン		①		②		③		④	
	マクロ	マイクロ	マクロ	マイクロ	マクロ	マイクロ	マクロ	マイクロ	マクロ	マイクロ
再現率	25.23	59.39	25.10	73.66	22.93	55.40	30.02	69.12	29.21	66.39
適合率	44.61	73.22	52.81	37.57	36.97	47.69	31.36	38.12	37.82	48.15
F-measure	**28.77**	**65.59**	19.44	49.76	24.98	51.25	25.04	49.14	29.46	55.82

表４：データ量における精度変化

データ量	1（ベースライン）		1/2		1/4	
	マクロ	マイクロ	マクロ	マイクロ	マクロ	マイクロ
再現	25.23	59.39	22.84	55.92	17.15	48.73
適合	44.61	73.22	46.35	72.91	32.56	70.96
F-値	**28.77**	**65.59**	27.04	63.29	20.80	57.78

表５：素性『TITLE』を追加しての実験

	ベースライン			素性『TITLE』追加		
	再現	適合	F-値	再現	適合	F-値
提案手法	6.83	24.14	10.65	14.15	38.16	**20.64**
解くべき問題	21.20	36.36	26.79	24.03	38.64	**29.63**
マクロ	14.02	30.25	19.16	19.09	38.4	**25.13**
マイクロ	15.16	33.18	20.81	19.88	38.49	**26.22**

た状態で Yamcha のベースとなる SVM の素性に関するパラメーターを変更した場合の実験を行った。さらに前述した『論文要約データのタイトル部分に出現することが多い』という性質を利用して、タイトル部分にのみ機械学習を用いる実験も行った。

この実験で用いる素性として、ベースラインのものに素性『TITLE』を追加した。素性『TITLE』は論文要約データ中のタイトル部分のみに付与し、タイトル以外にはダミーの素性『＊』を付与する。その結果を表５に示す。

SVM のパラメーターを変更した場合の実験では、素性としてベースラインの素性に素性『TITLE』を追加したものを用い、Yamcha のオプションを、

① FEATURE = F:-3..3:0.. T:-3..-1

② FEATURE = F:-4..4:0.. T:-4..-1

③ FEATURE = F:-5..5:0.. T:-5..-1

の順に変更して行うものとする。その結果を表６に示す。

タイトル部分のみを使用する実験では、論文要約データからタイトル部分のみを取り出し実験を行うが、この実験では素性『TITLE』は追加しなかった。これはタイトル部分だけを実験に使用する場合、『タイトル部分である』を意味する素性『TITLE』は必要ないからである。この実験では精度を求めるテストデータもタイトル部分のみを利用した。その結果を表７に示す。

表５を見ると、ベースラインに比べ、双方の分類共に精度は向上し、素性『TITLE』が有効に機能していることが分かる。

しかし「提案手法」「解くべき問題」はタイトルだけでなく本文中にも出現する。今後は本文中に出現するものに対して有効な素性を模索する必要がある。表６より、「提案手法」は③が、「解くべき問題」は①が最も良いことが分かる。

表７より、論文要約データのタイトル部分のみを用いた場合、ベースラインと比べ２倍以上の精度が得られた。タイトル部分のみを用いたため、集中的に学習が行えたことで、テストについても効率的にタグ付けできたと考えられる。

現段階では「正解の範囲」を拡張することにより、より精度を向上させることは可能であると考えられる。例えば、システムが「会話データ」と抽出したが、正解は「会話データの分析」であった場合、この「会話データ」も正解にするということである。しかし「正解の範囲」をどの程度取るかが問題となってくる。あまりに範囲が狭いと精度の向上は望めず、逆に範囲を広く取ると曖昧なものまで正解としてしまう。よって「正解の範囲」の決定が今後の課題の一つと言えるだろう。

３．２．２　「精度表現」

「精度表現」は通常「%」表記である。そこで正規表現パターンを作成して、適用することにより精度向上を図った。「（数字）%以上」「約（数字）%」といった正規表現パターンを人手で作成して「精度表現」を抽出した。

この実験では機械学習を用いるのではなく、正規表現パターンにより重要表現を抽出する。抽出後は他の実験と同

表６：SVM のパラメーターを変更した場合の実験

	パラメータ変更無し			①			②			③		
	再現率	適合率	F-値	再現率	適合率	F-値	再現率	適合率	F-値	再現率	適合率	F-値
提案手法	14.15	38.16	20.64	15.12	44.29	**22.55**	13.66	47.46	21.21	12.68	52	20.39
解くべき問題	24.03	38.64	29.63	24.38	40.12	30.33	23.32	42.86	30.21	24.38	43.67	**31.29**
マクロ	19.09	38.40	25.13	19.75	42.20	**26.44**	18.49	45.16	25.71	18.53	47.84	25.84
マイクロ	19.88	38.49	26.22	20.49	41.32	**27.40**	19.26	44.13	26.82	19.47	45.67	27.30

表７：タイトル部分のみでの実験

	出現総数	タイトル部分のみ		
		再現率	適合率	F-値
提案手法	84	41.67	57.38	48.28
解くべき問題	175	41.71	49.32	45.20
マクロ		41.69	53.35	46.74
マイクロ		41.70	51.67	46.15

様に精度を求める。

「精度表現」における実験結果を表８に示す。

表８：正規表現パターンでの実験

	再現率	適合率	F-値
ベースライン	11.11	100	20
正規表現	100	52.94	**69.23**

実験結果によると、ベースラインに比べ F-measure の値は 20%から 69.23%と大きく精度が向上した。しかし「%」は「精度表現」ではなく割合を表現する場合があり、割合を意味する「%」を「精度表現」としてしまう誤りがあった。割合を意味する「%」と「精度表現」を意味する「%」の違いを調べる必要がある。

４　おわりに

本稿では論文要約文集から重要情報を自動抽出する手法について述べた。ベースラインでは重要情報表現によって精度の差がかなりついたが、3.2 節の手法を用いることにより精度の低かった重要情報表現についても精度を向上させることができた。そして全分類のベースラインでの F-measure のマクロ・マイクロ平均は、28.77%、65.59%から 36.06%、67.12%に向上した。

今後の課題として、重要情報表現「組織・人名」については CRL 固有表現データと論文要約データを併用しデータ量（出現総数）を増やすことを検討する。「分かったこと」「数量的成果」については単語単位での抽出ではなく、文抽出の研究方法を利用し、文単位で抽出することを検討したい。それにより全重要表現について精度の向上を試みたい。本研究には重要情報表現のタグ付けの基準が不明瞭である、また、正しくタグ付けされていない箇所があるという問題がある。これらの問題も解決していきたい。

参考文献

［1］ Tomohiro Mitsumori, Sevrani Fation, Masaki Murata, Kouichi Doi and Hirohumi Doi ： Gene/protain recognition using Support Vector Machine after dictionary matching, BioCreative Workshop, 2004

［2］ 上島 豊, 佐々木 明, 森林 健悟, 井原 均, 村田 真樹, 白土 保, 井佐原 均：論文からデータベースを構築するための言語情報技術を活用したソフトウェアの開発、第６回光量子科学研究シンポジウム、2004

［3］ 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

［4］ 中野桂吾、平井有三：日本語固有表現抽出における文節情報の利用、情報処理学会論文誌、Vol.45、No.3、pp.934-941、2004

［5］ 工藤拓：汎用テキストチャンカー『Yamcha』version 0.30 使用説明書、奈良先端科学技術大学院大学、2004

［6］ NACSIS, NTCIR Workshop 1, Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term **Recognition, 1999**

［7］ 高村大也、松本祐治：独立成分分析を用いた文書分類 -SVM のための素性空間再構成-、情報処理学会研究報告、2001-NL-143、pp.17-24、May 2001