論文データからの重要情報の抽出と可視化

Extracting and visualizing important information from article abstracts

村田 真樹[*1] Masaki Murata
Stijn De Saeger[*1] Stijn De Saeger
橋本 力[*1] Chikara Hashimoto
風間 淳一[*1] Jun'ichi Kazama
山田 一郎[*1] Ichiro Yamada
黒田 航[*1] Kou Kuroda
馬 青[*2],[*1] Qing Ma
相澤 彰子[*3] Akiko Aizawa
鳥澤 健太郎[*1] Kentaro Torisawa

[*1]独立行政法人 情報通信研究機構 National Institute of Information and Communications Technology
[*2]龍谷大学 Ryukoku University
[*3]国立情報学研究所 National Institute of Informatics

The aim of the study was to find ways of extracting important information from natural language processing article abstracts. While many kinds of information extraction, such as from newspapers, web materials, or biology article abstracts are available,, few papers on extracting natural language processing information exist. It is hoped that this study will be useful to natural language processing researchers. Four categories were defined that contain important expressions for natural language processing article abstracts. A process of extracting these expressions was developed by using machine learning methods which extracted these expressions with an F-measure of 0.80. After considering partially correct expressions to be correct, the F-measure increased to 0.85. Afterwards, various kinds of visualization tools were prepared using extracted expressions. These tools can display the abstract of a paper with extracted important expressions highlighted in color or indicated in rows, and make a graph indicating the distribution and trends of papers including important expressions for each category.

表 1: 自動構築したサーベイの例

	精度	分野	言語	...
論文 1	91%	構文解析	日本語	...
論文 2	98%	形態素解析	日本語	...
論文 3	95%	形態素解析	英語	...
...	...	...	...	...
論文 33	58%	情報検索	日本語	...
...	...	...	...	...

1. はじめに

自然言語処理の論文アブストラクトから重要な情報を取り出す研究を行った．新聞や Web 文書や生物医学文献から情報抽出の研究はすでに多くなされているが，自然言語処理に関する情報を取り出す研究はあまりない [村田 07, 菊井 06, Murata 06, 近藤 07]．本研究は自然言語処理の研究者に役立つことを希望する．

教師あり機械学習を用いて自然言語処理の論文アブストラクトから重要な情報を自動的に抽出する方法を構築した．重要な情報を抽出するために教師データとなるタグ付けデータを作成し，それを用いて教師あり機械学習により重要な表現を抽出した．

抽出した表現は様々な目的で役立つ．関連する論文を検索するためのキーワードとして利用できる．自然言語処理のサーベイを自動的に構築するためにも利用できる．

重要な表現を論文アブストラクトから取り出し，行に論文のタイトルを含み，列に重要な表現の分類を含む表を作成した．表はサーベイに利用できる．そのような表の例を表 1 に示す．この場合,「精度」「分野」「言語」を重要な表現の分類として用いた．表はどのような課題でどのくらいの精度が得られるかを示している．次節に重要な表現の定義と重要な表現の取り出し方について説明する．

実際に，われわれの方法により取り出した重要な表現を使ってそのような表をユーザに示す種々のツールを構築した．これらのツールは，重要な表現を色付けして強調した論文のアブストラクトを表示できる．各列に抽出した重要な情報を含む表を表示できる．各分類の重要な情報を含む論文の分布や傾向を示すグラフを作成できる．これらのツールを 3. に示す．

関連研究として，Mitsumori らは機械学習を用いて生物医学分野において重要な表現であるタンパク質名を論文から取り出している [Mitsumori 04]．Sasaki らはパターンを記述した人手で作成した規則を利用することで物理学の論文から物理学で重要な表現 (例：原子記号や電子配置を示す記号）を取り出した．自然言語処理の論文から情報を取り出す研究としては文献 [村田 07, 近藤 07] があげられるが，これらはタイトルのみから情報を得ている．これに比べて本研究はアブストラクトからも情報を得る．

2. 重要な情報の抽出

自然言語処理の研究分野において，次の 4 個の重要な情報の分類を作成した．

1. 精度表現 — 精度を示す表現．(例：「97%」)
2. 主要な分野 — 自然言語処理における分野 (例：「機械翻訳」)
3. 言語名 — その論文が扱っている言語 (例：「日本語」)
4. 組織・人名 — 組織の名称や人名 (例：「ICOT」「ATR」)

連絡先: 村田 真樹，独立行政法人 情報通信研究機構知識創成コミュニケーション研究センター言語基盤グループ，〒 619-0289 京都府相楽郡精華町光台 3-5，TEL: 0774-98-6833, FAX: 0774-98-6961, murata@nict.go.jp.

タイトル:
日本語表層表現を手がかりとした名詞の指示性と数の推定

学会:
情報処理学会研究報告. 自然言語処理研究会報告

発表年:
1993

アブストラクト:
日本語[言語名] を 英語[言語名] に 翻訳[主要な分野] する時には,日本語[言語名] にはないが 英語[言語名] では必要な冠詞や数の問題に直面する. この難しい問題を解決するために, われわれは文章における名詞の指示性と数をそれぞれ三種類に分類した. この論文では, 名詞の指示性と数がその名詞の現れる文中の言葉によりかなりの程度推定できることを示した. その推定のための規則はエキスパートシステムの書き換え規則に類する形で, 文法書などから得られる知識をもとに経験的に作成した. これらの規則を作るのに利用したテキストでの正解率は, 指示性で 85.5%[精度表現], 数で 89.0%[精度表現] であった. 規則を作るのに利用していないいくつかのテキストでの正解率は平均して指示性で 68.9%[精度表現], 数で 85.6%[精度表現] という結果となった.

図 1: 可視化した論文の例

表 2: 実験結果

	完全一致			部分一致		
分類	再現率	適合率	F 値	再現率	適合率	F 値
精度表現	0.73	0.73	0.73	0.80	0.80	0.80
主要な分野	0.78	0.79	0.79	0.85	0.86	0.85
言語名	0.93	0.97	0.95	0.94	0.97	0.95
組織・人名	0.62	0.84	0.71	0.68	0.91	0.78
平均	0.77	0.83	0.80	0.83	0.89	0.85

分類の定義は重要である．これらの定義は自然言語処理の論文アブストラクトを注意深く考察して構築した．例えば,「言語名」の分類は，自然言語処理のための特別な分類であり，他の分野での情報の取り出しではあまり用いられないと思われる．

「精度表現」「主要な分野」「言語名」「組織･人名」の取り出しには，教師あり機械学習を利用した．YamCha を利用した．YamCha はサポートベクターマシン [Cristianini 00, Kudoh 00b, Kudoh 00a, 中野 04, Mitsumori 04] を機械学習法として用いて，フレーズの取り出しができる．われわれはタグ付けの仕方に IOB2[Kudoh 00a, 中野 04] を用いた．解析方向は文末から文頭とした． 2 次の多項式カーネルを用いパラメータ C には 1 を用いた．多値分類のために one versus rest 法を用いた．

機械学習のために以下の素性を用いた．

1. 解析対象の単語とその前方 3 単語と後方 3 単語とそれらの品詞と文字種 (平仮名か片仮名など).
2. 解析対象の単語の文節の主辞と種類の情報.

619 個のアブストラクトに対してタグ付けした．これらのアブストラクトは情報処理学会自然言語処理研究会で発表された論文から取り出した．このデータで 5 分割のクロスバリデーションで評価した.「精度表現」と「組織・人名」の分類については，このデータだけでは良好な精度が得られなかったため，1663 個の追加のアブストラクトも学習データに加えて利用した．この追加には情報処理学会の自然言語処理研究会と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を利用した．

実験結果を表 2 に示す.「完全一致」は，システムの出力があらかじめ作成した正解と完全に一致したときのみ正解とする.「部分一致」は，正解と一部でも重なった出力を正解との重なり具合も考慮して精度を求めたものである.「部分一致」の場合は，以下の式を使ってシステムの出力の精度を求めた.

$$num = \frac{2 \times agreement\_length}{correct\_length + answer\_length}, \qquad (1)$$

ただし， $agreement\_length$ は正解とシステムの出力の重なった文字数であり， $answer\_length$ はシステムの出力の文字数で，$correct\_length$ は正解の表現の文字数である．上記式の num の値をシステムの正解数として精度を計算した．表の平均は，4 個の分類での精度の平均である．

それぞれの分類とも，F 値は，部分一致の評価で 0.8 から 0.9 程度と高い．このことから，これらの分類におけるデータは自動で抽出できることがわかった．われわれの方法でこれらの分類に関係する自然言語処理のサーベイを自動で構築できることがわかる．

3. 可視化

次に，われわれの方法で自然言語処理において重要な表現を取り出し，重要な情報を表示する可視化ツールを構築した.「精度表現」「主要な分野」「言語名」「組織・人名」を抽出における F 値が高かったのでツールではこれらの分類を用いた.

3.1 論文のアブストラクトの可視化

我々は取り出した重要表現を色付けして強調表示して論文のアブストラクトを表示するツールを構築した．可視化したアブストラクト [村田 93] の例を表 1 に示す．抽出した重要な表現は図においてそれらの分類を示すタグをつけた下線が引かれている．[精度表現], [主要な分野], [言語名], [組織・人名] は，それぞれ「精度表現」「主要な分野」「言語名」「組織・人名」を指す．われわれの実際のシステムでは，重要な表現は分類ごとに異なる色付けをして表示される．この可視化されたアブストラクトでは，論文で強調表示されている精度表現を見ることで論文で記述されている技術の精度がどのくらいであるかを認識するのに役立つ．同様に主要な分野として強調表示され

表 3: 可視化した表の例

タイトル	年	精度表現	主要な分野	言語名	組織・人名
日本語の関係節における主辞の省略の解析	1997	９０%	意味，解析，語彙，認識	日本語	
保守性を考慮した日本語形態素解析システム	1997	９１．９%	解析，形態素解析	日本語	
並列 HPSG パーザーに向けて	1997		パーザー，辞書，素性		鳥澤
受け身/使役文の能動文への変換における機械学習を用いた格助詞の変換/受け身/使役文の能動文への変換における機械学習を用いた格助詞の変換	2002	８９．０６%，８９．５５%	素性		
Support Vector Machine を用いた決定性上昇型構文解析	2002	８８．２／８９．０%	コーパス，解析，解析木，語彙，構文解析，素性，文脈	英語	
SVM を用いた学習型質問応答システム SAIQA-II	2004	約５５%	質問応答，抽出	日本語	
意見文からの評判情報抽出に基づく自然言語検索	2006	９１．９%	クエリ，抽出		
GLR をベースにした自然言語処理用 MSLR パーザの改良	2007		パーザ，構文解析	日本語	東工大
統計的特徴を利用した機能語の自動認定実験	2007		辞書，翻訳	イタリア語，インドネシア語，スペイン語，フランス語，マレー語，ロシア語，英語，日本語	NHK
特許情報処理を指向したテストコレクションの構築：情報検索と自然言語処理の融合を目指して	2008		抽出，翻訳	日英	ＮＴＣＩＲワークショップ
機械翻訳最新事情：(上) 統計的機械翻訳入門	2008		機械翻訳，言語翻訳，翻訳	アラビア語，英語	

ている表現を見ることでこの論文の分野も即座に認識できる．このツールは，自然言語処理の研究者が自然言語処理のアブストラクトを読む際に役立つ．

3.2 抽出した重要情報の表による可視化

それぞれの列に抽出した重要な表現を含む表においてアブストラクトを表示する可視化ツールを構築した．可視化された表の例を表 3 に示す．表を見ることで多くの論文の特徴を一度にまとめて理解することができる．例えば，表の最後の論文はアラビア語を扱っていることがわかる．また 6 個目の論文から質問応答システムはだいたい 5 割程度の精度であることがわかる．この表は，多くの論文の状況や特徴を把握するのに役立つ．

3.3 抽出した情報のグラフによる可視化

主要な分野と言語名の分類の重要情報を含む論文の個数の分布を示すグラフを作る可視化ツールを構築した．図 2 と図 3 にこのツールで構築したグラフを示す．この分析には『NII-ELS 抄録データ 2009 年 01 月版 (国立情報学研究所)』を利用した．これはおよそ 700 万件の抄録データを持つ．このデータのうち，タイトル，アブストラクト，学会名など，抄録データのどこかに「自然言語」という文字列が含まれるデータを抜き出した．これは 3,928 件の抄録データであった．本節での実験はこのデータを用いて行ったものである．図 2 は主要な分野から見た論文の分布ををを示す．表から，それぞれの主要な分野にいくつの論文を持つかをはっきり見ることができる．最も大きな主要な分野は「抽出」であり，次に大きな主要な分野は「解析」であることがわかる．図 3 は言語名の分野の分布を示す．表から，それぞれの言語を扱う論文の数を認識できる．日本語が最も多く扱われ，次に英語，その次に中国が多く扱われていることがわかる．

最後にそれぞれの分類の重要表現を含む論文の傾向を示すグラフを作る可視化ツールを構築した．言語の分類を例にしたグラフを図 4 に示す．図から 1993 年から 1997 年に日本語に関する研究が多かったことがわかる．また，英語は 1990 年代からも多かったことがわかる．中国語の研究は 2000 年あたりから急に盛んになったこともわかる．図から，各言語を扱う論文の傾向を理解できる．

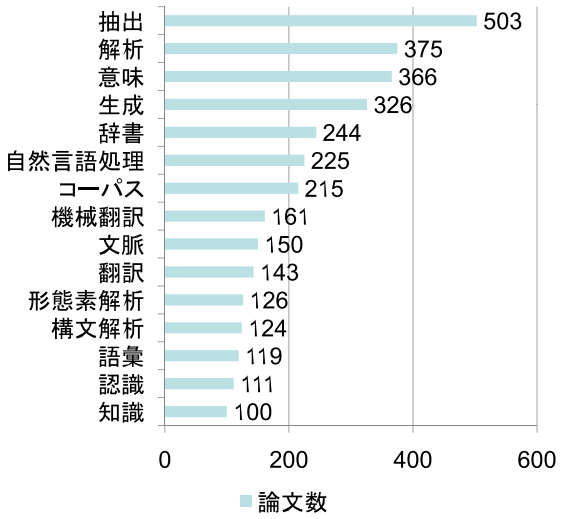

図 2: 主要な分野における論文の分布

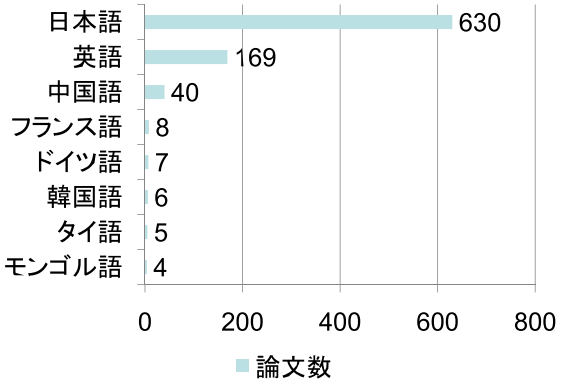

図 3: 言語名の分類における論文の分布

4. おわりに

教師あり機械学習を用いて「精度表現」「主要な分野」「言語名」「組織・人名」の 4 つの分類に属する重要表現を取り出す方法を構築した．われわれはこれらの表現を 0.80 の F 値で取り出した．部分的に正解した表現も正解とした場合では 0.85 の F 値を得た．これは，われわれの方法でこれらの分類の重要表現を自動で取り出すことができることを意味する．

次に抽出した重要な表現を用いた種々のの可視化ツールを構築した．構築したツールは，重要な表現を色分けして強調表示してアブストラクトを表示できる．また，重要な表現をそれぞれの列に含む表の形に複数のアブストラクトを整理して表示できる．また，重要な表現を含む論文の分布や傾向を示すグラフを表示できる．これらのツールは自然言語処理の論文調査に役立つ．

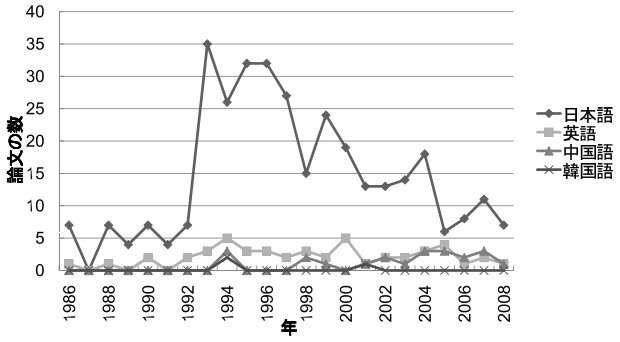

図 4: 言語名の分類での傾向

参考文献

[Cristianini 00] Cristianini, N. and Shawe-Taylor, J.: *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press (2000)

[菊井 06] 菊井 真, 村田 真樹, 馬 青：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, 言語処理学会第 12 回年次大会 (2006)

[近藤 07] 近藤 友樹, 難波 英嗣, 奥村 学, 新森 昭宏, 谷川 英和, 鈴木 泰山：論文データベースからの研究動向情報の抽出, 言語処理学会第 13 回年次大会, pp. 470–473 (2007)

[Kudoh 00a] Kudoh, T. and Matsumoto, Y.: Use of Support Vector Learning for Chunk Identification, *CoNLL-2000*, pp. 142–144 (2000)

[Kudoh 00b] Kudoh, T.: TinySVM: Support Vector Machines, http://cl.aist-nara.ac.jp/ taku-ku// software/TinySVM/ index.html (2000)

[Mitsumori 04] Mitsumori, T., Fation, S., Murata, M., Doi, K., and Doi, H.: Gene/protain recognition using Support Vector Machine after dictionary matching, *BioCreative Workshop: Critical Assessment for Information Extraction in Biology (BioCreative 2004)* (2004)

[村田 93] 村田 真樹, 黒橋 禎夫, 長尾 真：日本語表層表現を手がかりとした名詞の指示性と数の推定, 情報処理学会自然言語処理研究会 93-NL-95, pp. 33–40 (1993)

[Murata 06] Murata, M., Kikui, M., Ma, Q., Kanamaru, T., and Isahara, H.: Extracting important information from natural language processing article abstracts and visualizing it, *Proceedings of Joint 3rd International Conference on Soft Computing and Intelligent Systems and 7th International Symposium on advanced Intelligent Systems*, pp. 112–117 (2006)

[村田 07] 村田真樹, 一井康二, 馬青, 白土保, 金丸敏幸, 井佐原均：過去 10 年間の言語処理学会論文誌・年次大会発表における研究動向調査 (2007), 言語処理学会ホームページ (http://www.nak.ics.keio.ac.jp/NLP/trend-survey.html)

[中野 04] 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 (2004)