2009－NL－190 (18)
2009／3／26

系列パターンを利用した決定木による自然言語における選択ルール獲得

古宮嘉那子[†] 柴原一友[†] 但馬康宏[†] 藤本浩司[††] 小谷善行[†]

東京農工大学[†] テンソル・コンサルティング株式会社[††]

{komiya, k-shiba}@fairy.ei.tuat.ac.jp, {ytajima , kotani}@cc.tuat.ac.jp

koji.fujimoto@tensor.co.jp[†]

Abstract

自然言語文から抜き出した系列パターンの有無を要因に，決定木学習を用いて自然言語における選択ルールを生成した．人が考えた要因を入力にして決定木学習を行った場合には劣るものの，「結構」の意味判定の問題と，話者の性別判定の選択ルール生成において，両者ともに3.25ポイントの差まで迫る正解率を示した．意味に関しては，単純に系列パターンを要因にした場合が最もよく，性別に関してはNgramを決定木学習の最中に要因の候補として取得する場合が最もよかった．生成されたルールを見ると，決定木上部では，人が与えた要因による決定木の上部に現れる要因が多く選ばれていることが分かった．

Acquisition of a Set of Determination Rules in Natural Language Using Sequential Patterns

Kanako Komiya[†], Kazutomo Shibahara[†], Yasuhiro Tajima[†], Koji Fujimoto[††] and Yoshiyuki Kotani[†]

Tokyo University of Agriculture and Technology[†]
Tensor Consulting Co. Ltd.[††]

Abstract

The presences or absences of sequential patterns that are extracted from sentences are used in order to acquire a set of determination rules in natural language using decision tree learning. The differences of accuracies of the rules were only 3.25 points for two problems: the determination of the meaning of "kekkou" and the determination of the gender of a speaker, though they were less than those of the rules using the features humans selected.

Most of the features in the upper part of the generated trees are also appeared in the upper part of the trees using the features humans selected.

1． はじめに

自然言語処理において，決定木学習を用いて，多義語の意味や，敬語の使用方法などを選択する研究が行われている[1-3]．特に筆者は，決定木学習を用いて，多義語の意味や，敬語の使用方法などを適切に使用するルールを得ることを目的とした研究を行ってきた[2, 3]が，その際，決定木学習に用いられる学習データの要因や特徴量は問題の特徴にあわせて人手で与える必要があった．

一方，バスケット分析の分野で用いられる相関ルールでは，データから直接，要因も同時に抽出できる．自然言語処理の分野でも，相関ルールのApriori Algorithm[4]を，順序を考慮した系列パターンに拡張したアルゴリズム，PrefixSpan[5]によって得られた系列パターンを機械学習の入力に用いた研究がなされている[6-9]．

本論文では，PrefixSpanによって自然言語文から抜き出した系列パターンの有無を要因に，決定木学習を用いて自然言語における選択ルールを生成した．

2. 属性選択システム

選択ルールの生成とルールの正解率の測定のために，属性選択システムを作成した．属性選択システムは，用例により自動的にルールを作成して，それを基に適切な属性を選択するものであり，どのような原因でどのような用法をするかなど，自然言語の選択ルールを明示的に分析できる利点を持つ．属性選択システムは，決定木作成フェーズと実行フェーズからなっている．決定木作成フェーズでは，属性を選択するための要因と，そのときの属性からなる学習データによって決定木学習を行うことで属性選択ルールを作成する．また，実行フェーズでは，テストデータの要因に対して，決定木作成フェーズで得られた属性選択ルールを適用し，得られた結果の正否を見ることで，正解率を測定する．(図 1)

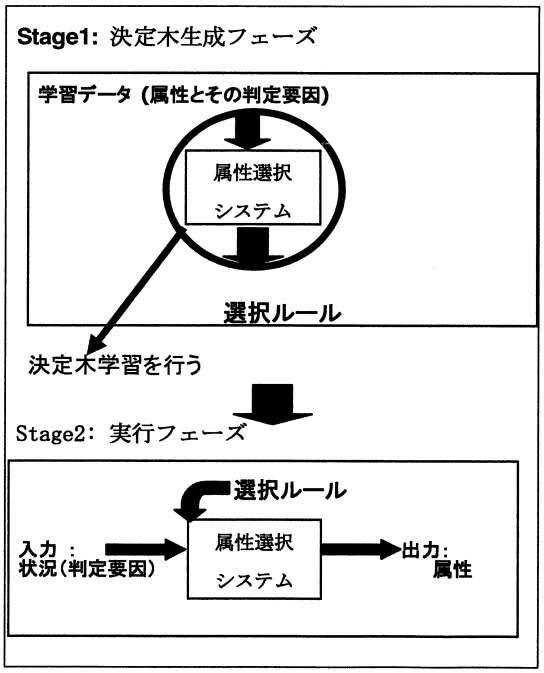

図 1　属性選択システム

本研究では，属性選択システムの入力となる要因に系列パターンを利用し，人手で与えた知識を要因とした場合と比較する．決定木の生成には，C4.5[10]を用い，決定木は二分木を生成した．

3. 要因としての系列パターン

本研究では，系列パターンの有無を要因とする．系列パターンとは，順序を考慮したパターンである．Ngram もこれに属するが，系列パターンは，要素同士が隣接していなくてもよい．例えば，(a, b, c)から得られる系列パターンは，{(a), (b), (c), (a, b), (a, c), (b, c), (a, b, c)}の 7 つとなる．本研究では，形態素をひとつの要素と考え，系列パターンを文から抽出し，この有無を決定木学習の要因として自然言語の選択ルールを作成した．

これらの系列パターンは，PrefixSpan によって抽出した．PrefixSpan とは，相関ルールの Apriori Algorithm を系列パターンに拡張したパターン列挙のアルゴリズムであり，数え上げるパターンの最低の頻出トランザクション数（サポートと呼ばれる）を設定することにより，効率よく主要なパターンを抽出する手法である．

同音異義語や違った活用の形態素を，別の要素として認識するために，PrefixSpan を行う前に，本研究では，文を形態素解析し，単語，読み，原形，品詞，活用，型の六種類の情報をセットにしてひとつの要素とした．(cf. 付録　図 2)

形態素解析には茶筌[11]を用いた．また，PrefixSpan には PrefixSpan-rel[12]を用いた．この際，右最大のパターンに限らず，全パターンを出力した．(a と a b がともにサポート以上なら，a と a b をともに抽出)

4. 意味/性別選択とその判定要因

本章では，実験に利用したデータとその要因について述べる．系列パターンを要因としたルール生成を行う際，「結構」の意味選択と，話者の性別判定の二つの問題について試した．意味選択にはインターネットから得られた「結構」(名詞または形容動詞語幹) を含む文 500 件を利用し，性別判定には小説の会話文 1230 件を利用した．この際，データの正解は人手で与えた．表 1 と表 2 に，それぞれ意味選択の問題と性別選択の問題の属性の種類とその件数およびパーセンテージを示す．

また，ルール作成のために，決定木学習を行う際の要因に，「結構」の意味選択については以下の 1)～3)の三通り，性別選択では以下の 1)～4)の四通りを試した．以下，それぞれについて述べる．

1) 人手で与えた要因
2) 単純な系列パターンと Ngram
3) 結果ごとの系列パターンと Ngram
4) 動的な系列パターンと Ngram

表 1　意味選択システムの属性の種類とその件数およびパーセンテージ

属性の種類	件数[件]	パーセンテージ　[%]
かまわない	289	57.80
いらない	111	22.22
すばらしい	100	20.00

表 2　性別選択システムの属性の種類とその件数およびパーセンテージ

属性の種類	件数[件]	パーセンテージ　[%]
男性	615	50.00
女性	615	50.00

4.1　人手で与えた要因

(1)結構の意味選択

「結構」の前後の形態素が文脈を形作っているという仮定に基づき，「結構」の含まれた文例を形態素解析し，「結構」の前後の形態素について，単語，読み，原形，品詞，活用，型の六種類の情報を要因とする．

例えば，「だったら結構だよ」という文章の場合は，「結構」の前後の形態素がそれぞれ
前々：「だっ」，前：「たら」，後：「だ」，後後：「よ」
であるため，図 3 の 24 の要因を用いる．(付録参照)

(2)話者の性別判定

文末の形態素と終助詞，代名詞が男言葉/女言葉に大きく影響しているという仮定に基づき，小説から抜き出した会話文 1230 件を形態素解析し，会話文の文末から一番目と二番目の形態素について，単語，読み，原形，品詞，活用，型の六種類の情報を要因とする．

例えば，「だったら結構だよ」という文章の場合は，会話文の文末から一番目と二番目の形態素がそれぞれ，文末からふたつめ：「だ」，　文末からひとつめ：「よ」であるため，図 4 の 12 の要因を用いることになる．(付録参照)

さらに，文例中の品詞が「名詞-代名詞-一般」または「助動詞-終助詞」である形態素の有無を要因に加える．要因の数は文例によって変化するが，本実験では，「僕」などの 45 の代名詞と，「わ」や「ね」などの 19 の終助詞を利用した．

4.2　単純な系列パターンと Ngram

全データから一括に PrefixSpan によって系列パターンを抽出する．(Ngram を含んだ) 系列パターン (今後単に系列パターンと表記) と Ngram のみの二種類について試した．

4.3 結果ごとの系列パターンと Ngram

学習データの結果 (「結構」の意味や話者の性別) 別にデータを分け，それぞれ PrefixSpan によって系列パターンを抽出する．このことによって，結果に直接関係したパターンが得られやすくなると考えた．

4.4　動的な系列パターンと Ngram

決定木学習の途中で，分類したデータの量によってサポートを変えながら動的に系列パターンを得た．このことによって，決定木上部では多く頻出する系列パターンを，下部では少量の系列パターンを利用することができると考えた．話者の性別選択の際だけに使用した．

5．　属性選択システムの実験

五分割交差検定によって正解率を見た．閾値は以下の二通りを利用し，最も正解率のよいものを採用した．

- エントロピー
- エントロピー×データ件数

属性選択システムは，葉の数により結果の数が異なるため，葉の数により正解率が変化する．そのため，できるだけ葉の数に偏りが出ないように実験を行った．PrefixSpan のサポートは，いくつか予備実験を行った結果，要因数 3000 前後のとき，正解率が高くなったため，常に要因数ができるだけ 3000 前後に近くなるように設定した．また，結果ごとの系列パターンと Ngram を得る際は，できるだけ結果ごとの要因数がデータ数の比に近くなるようにしたため，「結構」の意味選

択についてはそれぞれサポートに，「かまわない」を 8，「いらない」を 5，「すばらしい」を 7 に設定した．

また，動的な系列パターンと Ngram では，データ量のノード中のデータ件数（トランザクション数）の 1/100 をサポートとし，パターンの種数が多くなりすぎた場合には，サポートを 1 追加することで対応した．また，データ件数が 200 件未満の場合には，特例としてサポートを 2 とした．

6． 属性選択システムの結果と選択ルール

付録の表 3 と表 4 にそれぞれ，「結構」の意味選択の問題における試み別による正解率，葉の数，要因数の表を示す．結果ごとの系列パターンと Ngram を得る際は，全件に一括に PrefixSpan を行わず，五分割交差検定で利用する全体の 4/5 の学習データに対して行ったため，試行ごとに要因数が変化する．そのため，平均値を示した．

6.1 「結構」の意味選択ルール

表 3 から，「結構」の意味選択の問題では，Ngram よりも，系列パターンを利用したときの方が常に正解率が高いことが分かる．ルールを見てみると，全体から Ngram を抽出した際は，サポートを 2 まで下げたことで，二度現れる長い Ngram が要因として選ばれやすくなっていた．また，結果ごとに Ngram を抽出した際には，副詞「結構」の有無がルールの上部に現れていた．本研究の「結構」は名詞または形容動詞語幹のみを対象としているため，人手の要因の時には起こりようのない誤りである．

また，系列パターンの決定木に選ばれた要因は，「結構」を含んでおり，人手で与えた「結構」の前後の形態素情報と等しいものが多かった．ただし，品詞情報のみの入力がないため，「名詞-非自立-一般」（人手の場合のトップノードの要因）のような要因は抽出することができなかった．

また，「結構」の意味選択の問題では，結果ごとに抽出するときよりも一括の方が，正解率と可読性（葉の数が少ない方がよい）または可読性のみにおいて優れている．これは，学習データが 500 件と少量であり，分類クラスも 3 クラスであることから，結果ごとの学習データが少なくなり，特色が出にくいためであると考えられる．

最もよいルール（単純な系列パターン）に現れた要因の例を一部示す．

1. 結構ケッコウ結構名詞-形容動詞語幹なナだ助動詞特殊・ダ体言接続//. . . 記号-句点
2. しばらくシバラクしばらく副詞-助詞類接続
3. とてもトテモとても副詞-助詞類接続
4. 結構ケッコウ結構名詞-形容動詞語幹なナだ助動詞特殊・ダ体言接続//だダだ助動詞特殊・ダ基本形　（注：//は一つ以上の形態素を表す）

6.2 話者の性別選択ルール

話者の性別判定でも，決定木の上部に文末の形態素や，「僕」などの代名詞，「わ」などの終助詞など，人手で与えた要因と同じ要因が現れた．また，「けど」を使うと女性，「が」を使うと男性が多いなど，接続助詞の要因が多く選ばれ，人手で与えなかった特色も見ることができた．

しかし，ルールを比較してみると，人手で与えた場合にルールに現れる要因のうちには，サポート 10 または 7 では候補にあがらない形態素もあり，そのため，サポートの低い Ngram が，正解率が高い．

また，話者の性別選択の問題では，一括よりも，結果ごとに抽出するときの方が正解率と可読性において優れている．これは，学習データが 1230 件，分類クラスが 2 クラスであることから，結果ごとの学習データが比較的多くなり，特色が出るようになったためと考えられる．例外が結果ごとの Ngram の可読性であるが，サポートが 2 であるため，上部から細部にわたってルールを作っており，複雑な決定木になっている．

このようなサポートによるトレードオフを解決するために，動的な要因の取得を行ったところ，最も結果がよくなったと考えられる．

最もよいルール（動的な Ngram）に現れた要因の例を一部示す．

1. わワわ助詞-終助詞
2. のノの助詞-終助詞 よヨよ助詞-終助詞
3. けどケドけど助詞-接続助詞 ,,, 記号-読点
4. 僕ボク僕名詞-代名詞-一般

さらに，ふたつの問題において，特に系列パターンでは，句読点や「に」や「を」などの助詞が

よく現れた．これらは全般的によく現れる形態素であり，実際にはルールにはあまり関係ないことが示唆される．このため，系列パターンを要因とする際，サポートだけではなく，リフト率のように，単独使用数との比較から重要であることを測る指標を利用することによって改良できる可能性があると考えられる．

7． まとめ

自然言語文から PrefixSpan を用いて抜き出した系列パターンの有無を要因に，決定木学習を用いて自然言語における選択ルールを生成した．人が考えた要因を入力にして決定木学習を行った場合には劣るものの，「結構」の意味判定の問題と，話者の性別判定の選択ルール生成において，両者ともに 3.25 ポイントの差まで迫る正解率を示した．意味に関しては，単純に系列パターンを要因にした場合が最もよく，性別に関しては Ngram を決定木学習の最中に要因の候補として取得する場合が最もよかった．生成されたルールを見ると，決定木上部では，人が与えた要因による決定木の上部に現れる要因が多く選ばれていることが分かった．また，系列パターンと Ngram 両者において，句読点や「に」や「を」などの助詞がよく現れたため，単独使用数との比較から重要であることを測る指標を利用することによって改良できる可能性があると考えられる．

参考文献

[1] 木村 直樹, 松原 茂樹, 小川 泰弘, 稲垣 康善: 音声対訳コーパスからの日本語待遇表現生成規則の自動獲得．情報処理学会研究報告，2000-NL-148，pp. 37-43，（2000）．

[2] Kanako Komiya, Yasuhiro Tajima and Yoshiyuki Kotani, Generating a set of rules to determine the meanings of "kekkou", Proceedings of the 10th World Conference on Integrated Design and Process Technology (IDPT2007), pp. 202-209, (2007) .

[3] Kanako Komiya, Chikara Igarashi, Kazutomo Shibahara, Koji Fujimoto, Yasuhiro Tajima and Yoshiyuki Kotani, Generating a Set of Rules to Determine the Gender of a Speaker of a Japanese Sentence, WSEAS TRANSACTIONS on COMMUNICATIONS, pp. 112-121, (2009) .

[4] Agrawal, R. and Srikant, R.: Fast Algorithms for mining association rules, Proceedings of the 20th VLDB Conference, pp. 487-499 , （1994）.

[5] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal and Mei-Chun Hsu: PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth, Proc.of International Conference of Data Engineering (ICDE2001), pp. 215-224 (2001).

[6] Masahiro Terabe, Takashi Washio, Hiroshi Motoda, Osamu Katai and Tetsuo Sawaragi: Attribute Generation Based on Association Rules, Knowledge and Information Systems, 4: pp. 329-349, Springer-Verlag, London Ltd. , (2002) .

[7] 山本薫,工藤拓,坪井祐太,松本裕治：系列パターンマイニングによる対訳表現抽出，情報処理学会研究報告，2002-NL-149, pp.15-22, (2002) .

[8] 工藤拓，松本裕治：系列パターンマイニングを用いた有効な素性の組み合わせの発見，情報処理学会研究報告，2002-NL-153，pp. 147-154, (2003) .

[9] 磯崎秀樹，平尾努，鈴木潤：機械学習のための組み合わせ素性の選択基準について，情報処理学会研究報告，2003-NL-158 , pp. 63-68 , (2003) .

[10] J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers Inc, (1993).

[11] MATSUMOTO, Y. et al: Japanese Morphological Analysis System ChaSen version 2.2.1, http://chasen.aist-nara.ac.jp/chasen/doc/chasen-2.2.1.pdf, (2000) .

[12] 浅原 正幸, PrefixSpanrel 系列パターンマイニングツール, http://prefixspanrel.sourceforge.jp/, （2007）.

付録

そうソウそう副詞-助詞類接続　だダだ助動詞特殊・ダ基本形　ねネね助詞-終助詞

図 2　「そうだね」の書式のデータ

単語，読み，原形，　品詞，　活用，　型
前前：だっ，ダッ，だ，助動詞，特殊・ダ，　連用タ接続
前　：たら，タラ，た，助動詞，特殊・タ，仮定形
後　：だ，　ダ，　だ，助動詞，特殊・ダ，基本形
後後：よ，　ヨ，　よ，助詞-終助詞，なし，なし

図 3　「結構」の意味選択における「だったら結構だよ」の要因

単語，読み，原形，　品詞，　活用，　型
文末からふたつめ　：だ，　ダ，　だ，助動詞，特殊・ダ，基本形
文末からひとつめ　：よ，　ヨ，　よ，助詞-終助詞，なし，なし

図 4　性別判定における「だったら結構だよ」の要因

表 3　「結構」の意味選択の問題における試み別による正解率，葉の数，要因数

試み	種類	サポート	正解率[%]	葉の数	要因数（平均）
人手	人手による	なし	85.60	53	921
単純	系列パターン	12	82.00	18	2987
単純	Ngram	2	79.40	44.6	3523
結果ごと	系列パターン	8, 5, 7	82.00	29.6	3117.4
結果ごと	Ngram	2	76.60	48.4	2227.4

表 4　話者の性別選択の問題における試み別による正解率，葉の数，要因数

試み	種類	サポート	正解率[%]	葉の数	要因数（平均）
人手	人手による	なし	69.27	125	1342
単純	系列パターン	10	61.71	18.4	2997
単純	Ngram	3	64.07	121.2	1867
結果ごと	系列パターン	7	62.20	38.2	2763
結果ごと	Ngram	2	64.47	205.4	2785.6
Pre 動	系列パターン	随時変更	61.30	152.6	随時変化
Pre 動	Ngram	随時変更	66.02	88	随時変化