E-006

文章を整理するための表自動生成手法
Automatic Table Creating Method for Organizing Sentences

西口 駿祐† 芋野 美紗子† 土屋 誠司‡ 渡部 広一‡
Shunsuke Nishiguchi Misako Imono Seiji Tsuchiya Hirokazu Watabe

1. はじめに

近年，急速な技術の発達により，私たちの生活の中には膨大な情報が存在している．そのため，そこから求める情報を取得するには多大な時間と労力が必要である．その解決方法として，本稿では表を用いて文章を整理する手法を提案する．表は，人名や地域などを見出しとする項目と，各項目に分類された語によって表現される．そのため表の生成には適切な項目の取得が必要であると考える．そこで，文章に適した項目を自動で生成することで，表を用いて文章を整理するシステム(以後，表生成システム)を構築する．

2. 研究概要

表生成システムはまず入力した文章から一文ごとに自立語を取得する．自立語から見出しとなる項目を生成する．その後，生成した項目に自立語を格納する．すべての自立語から項目を決定し，自立語を格納することで表を生成する（表 1）．表生成後，格納されている自立語の移動や項目の統合を行い，表を再生成する．

表 1　文章より自動生成した表

時刻	人物	場所	スポーツ	ゲーム	動詞
昨夜	私	球場	野球		した
先週	男	会社			刺された
去年		会社		麻雀	開かれた
一昨年	彼		サッカー		勝利した

3. 表生成システム

表生成システムの流れを図 1 に示す．

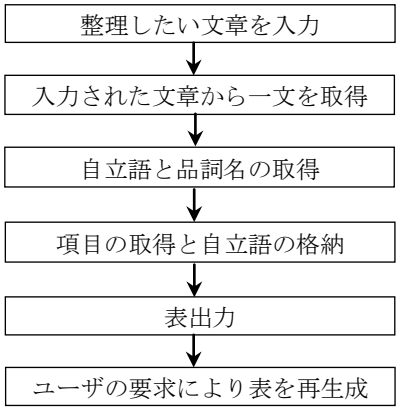

図 1　表生成システムの流れ

3.1. 入力された文章から自立語と品詞名の取得

まずユーザが整理したい文章を入力する．入力する文章は分野を問わず，複数文の入力が可能である．入力された文章を一文ごとに茶筅[1]を用いて形態素解析を行い，自立語とその品詞名を取得する．シソーラス[2]には一般名詞の意味的用法を表す 2710 個の意味属性(ノード)の上位－下位関係，全体－部分関係が木構造で示されており，約 13 万語が登録されている．

文の意味は，主に名詞と動詞，形容詞から理解が可能であると考え，表生成システムでは品詞が「名詞」，「動詞」，「形容詞」と判断された自立語のみを取得する．このとき，「携帯-電話」や「電子-辞書」など名詞と判断される自立語が連続していた場合は，それらを複合語と判断して一つの名詞とする．また，「大きい山」や「美しい顔」など形容詞の次の自立語が名詞または動詞と判断された場合，その形容詞は修飾語であると考え，複合語として判断する．複合語と判断された場合は，語尾の自立語の品詞を複合語の品詞とする．たとえば「大きい山」という複合語の場合，「山」が名詞であるため，「大きい山」は名詞と判断する．

3.2. 項目の生成と自立語の格納

取得した自立語の品詞名が動詞であれば「動詞」項目を生成し，形容詞であれば「その他」項目を生成する．取得した品詞名が名詞の場合，その自立語を格納する項目を判断する．品詞が名詞の場合の項目決定の流れを図 2 に示す．

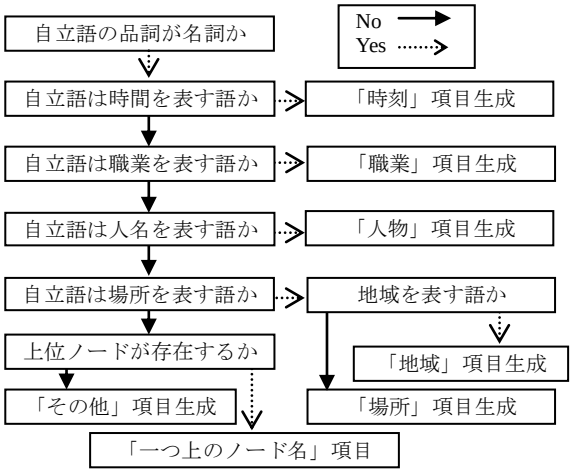

図 2　表の項目決定の流れ

図 2 より，自立語の品詞が名詞と判断された場合の項目生成方法を述べる．まず自立語が時間を表す語であるかを時間判断システム[3]を用いて判断する．時間判断システムとは入力された語が時間を表しているか判断し，時間を表

†同志社大学大学院工学研究科
Graduate School of Engineering, Doshisha University
‡同志社大学理工学部
Faculty of Science and Technology, Doshisha University

している場合はその語がいつ始まりいつ終わるかが出力されるシステムである．自立語が時間を表す語である場合は「時刻」項目を生成する．項目を生成しない場合は茶筌を用いて，自立語が職業を表す語であるか，または人名を表す語であるか判断する．自立語が職業を表す語である場合は「職業」項目を生成し，人名を表す語であれば「人物」項目を生成する．項目を生成しない場合は，場所判断システム[4]を用いて自立語が場所を表す語であるか判断する．場所判断システムとは入力された語が場所を表しているか判断し，そこに存在する人や物，そこで行う目的が出力されるシステムである．自立語が場所を表す語であると判断された場合は，茶筌を用いて地名を表す語であるか判断する．地名を表す語である場合は「地域」項目を，地域を表す語でない場合は「場所」項目を生成する．場所を表す語でない場合は，シソーラスを用いて上位ノードを項目名として取得する．上位ノードが取得できない場合は，「その他」項目を生成する．自立語から項目を生成した場合，項目に自立語を格納し，別の自立語の項目を判断する．システムが入力した文章のすべての自立語から項目を決定した場合は表を生成する．

4. 表の再生成

生成した表よりユーザの要求に対応した表に再生成する．

4.1. 項目または格納されている自立語の移動

生成した表より，ユーザが表の項目や格納されている自立語を指定した場合，指定した項目または自立語を表の最も左上に移動して表を再生成する．例を表 2 に示す．

表 1 より，ユーザが「会社」を指定した場合，「会社」を表の最も左上に移動して表を再生成する（表 2）．

表 2　ユーザの指定に対応し，再生成した表

場所	時刻	人物	スポーツ	ゲーム	動詞
会社	先週	男			刺された
会社	去年			麻雀	開かれた
球場	昨夜	私	野球		した
	一昨年	彼	サッカー		勝利した

表 2 より，会社を最も左上に移動することで，ユーザは会社で起きた文の取得が容易になると考える．

4.2. 項目の統合

ユーザが項目の統合を求める場合，シソーラスによりすべての項目の上位ノードを取得し，共通する上位ノードを持つ項目同士を統合する．その後，共通の上位ノードを項目として表を再生成する．例を表 3 に示す．表 3 は表 1 より，項目の統合を行い再生成した表である．「ゲーム」と「スポーツ」の項目は，上位ノードに共通して「娯楽」が存在するため，新たに「娯楽」項目として統合し，再生成した表である．

表 3　表 1 より，項目を統合して再生成した表

時刻	人物	場所	娯楽	動詞
昨夜	私	球場	野球	した
先週	男	会社		刺された
去年		会社	麻雀	開かれた
一昨年	彼		サッカー	勝利した

項目を統合することで，表が縮小化され，文章の整理が可能となると考える．

5. 表生成システムの評価

入力する文章は毎日新聞および朝日新聞のニュースサイトから，歴史，経済，スポーツ，生活，事件に関する文を 100 文ずつ取得した．取得した文から表生成システムを用いて表を生成し，3 人にアンケートを行い，評価した．評価内容は，生成された項目が正しいか，および自立語は適切な項目に格納されているについて評価を行った．評価は，3 人中 3 人が正解の場合は「○」，3 人中 2 人が正解の場合は「△」，正解と評価した物が 3 人中 2 人未満の場合は「×」とした．△は 3 人中 2 人が正解と判断した評価であるため，精度は△の割合に 2/3 倍した値に○の割合を加算した値を精度とする．評価結果を表 4,5 に示す．

表 4　生成された項目の精度

評価	割合
○	51.9%
△	25.4%
×	22.7%

表 5　格納された自立語の精度

評価	割合
○	47.3%
△	10.1%
×	42.6%

表 4 より，生成された項目が正しいかについての割合は○が 51.9%，△が 25.4%であった．よって△の割合に 2/3 倍した値 16.9%に○の割合を加算すると，精度は 68.8%を得た．

また，表 5 より，格納された自立語の割合は○が 47.3%，△が 10.1%であった．よって精度は 54.0%を得た．

6. おわりに

本稿では，膨大な文章から表を用いて整理を行うための表自動生成手法を提案した．そして表生成システムの評価は，自動で生成された項目の正しさと，格納されている自立語の正しさについて行った．その結果，項目の自動生成についての精度は 68.8%，格納されている自立語の精度は 54.0%を得ることができた．

謝辞

本研究の一部は，科学研究費補助金（若手研究（B）21700241）の補助を受けて行った．

参考文献

[1] ChaSen‥ 形態素解析器，奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座（松本研究室），http://chasen-legacy.sourceforge.jp/，(1997).

[2] NTT コミュニケーション科学研究所監修，「日本語語彙体系」，岩波書店，1997.

[3] 宮柳皓介，吉村枝里子，渡部広一，河岡司，“常識を持つコンピュータの実現に向けた常識的道具判断システムの構築”，FIT2008，E-054，2008.

[4] 奥村紀之，土屋誠司，渡部広一，河岡司，“概念間の関連度計算のための大規模概念ベースの構築”，自然言語処理，Vol.14，No.5，pp.41‐64，2007.