自然言語処理 137－9
音声言語情報処理 31－4
（2000．6．2）

日本語略語の自動復元

石井 直樹＊ 平石 智宣 延澤 志保 斎藤 博昭 中西 正和

慶應義塾大学 理工学研究科 計算機科学専攻
{naoki,tomonobu,shiho,hxs,czl}@nak.ics.keio.ac.jp

日本語略語を復元するシステムについて報告する。このシステムは、任意の日本語略語に対して、新聞記事コーパス中の語句および辞書中の語句のうちから、いくつかの復元規則を用いて、元の語になると考えられるものを順位を付けて出力するものである。復元規則として、「元の語が略語内の文字を全て、同じ順で含むこと」、「略語と元の語を構成する字種が等しいこと」、「元の語の文字数が略語を構成する字数の 4 倍以内であること」、「略語内の文字が元の語の中で不連続的に含まれていること」といったことを定めた。用いる復元規則の数を変えながら 404 の略語に対して実験を行い、7 割以上の確からしさで復元に成功した。

キーワード：略語，略語復元

Restoring Japanese Abbreviations

Naoki ISHII Tomonobu HIRAISHI Shiho NOBESAWA
Hiroaki SAITO Masakazu NAKANISHI

Department of Computer Science
Keio University
3-14-1, Hiyoshi, Kouhoku-ku, Yokohama 223-8522, Japan

This paper reports a system which restores Japanese abbreviations. This system accepts an abbreviation and hypothesizes its original phrases with likelihood. The system chooses the candidates from a newspaper corpus and three dictionaries using some restoration heuristics. The heuristics are: "the candidate word includes all the constituents of the abbreviation in the same order", "the abbreviation and the candidate use the same character set", "the number of characters of the candidate is less than four times of the number of the abbreviation characters", and "the characters of the abbreviation do not appear successively in the candidate." The system is tested against 404 abbreviations and over 70% of them are restored successfully. Effectiveness of the heuristics is also evaluated.

KeyWord: abbreviation, abbreviation restoring

＊現ソニー株式会社勤務

1 はじめに

近年、電子文書の普及により大量の文書を効率良く扱うことができるようになった。しかし、機械的な処理だけでは、同義語、照応、略語に対応できないという問題も生じている。

これらの中で同義語や照応は多くの研究がされているが、略語の研究は、特に日本語の場合、ほとんど未解決のままである。また、英語の略語においても、プログラム中での略語など対象を絞ることで復元を可能とする [1] など、一般的な略語についてはその効果的な解決法は、未だ見つかっていない。

そこで、本研究では復元という方針で略語の問題を解決することを考える。そして、日本語における略語の規則を考えることで、略語の復元をどの程度正確に行うことができるかを探る。また、有効な略語の規則を発見することにより、その精度を上げることを目標とする。

2 略語

2.1 略語の特徴

略語とは「語形の一部を省略して簡略にした語」(広辞苑) のことである。そのため、略語にはその元になる省略前の語が存在することになる。本来略語は有用なものであるが、文章内で利用されるとそのニュアンスは伝わるが元の語がわからないことがある。この原因としては、略語は正式な取り決めまたは規則の下で決まるものは少なく、口調などにより決定されることがあるなど曖昧な部分を多く含んでいることが考えられる。そのため、機械的に略語を生成、復元することは非常に困難である。

また、略語の定義にも曖昧な点がある。まず、略語にも段階があるということが挙げられる。例として、「慶應義塾大学」という語を考えると、その略語として「慶應大学」「慶應大」「慶大」と 3 通りを考えることができる。このような際に「慶應義塾大学」の略語としてはどの略語が適切であるかという問題がある。これらの略語間には、適切・不適切という差は存在しない。また、「慶應義塾大学」の略語 (「慶應大学」) の略語 (「慶大」) も存在し得ることになる。逆に「慶大」を復元する際には、どのレベルまで復元できればその語の復元に成功したことになるかという問題もある。また、「ダイヤ」という略語に関しては少なくとも「ダイヤモンド」「ダイヤグラム」と 2 通りの可能性が存在するという問題もある。

略語においては、単純に復元するというだけではなく、評価をする際の基準を明確にし、判定をする必要がある。また、常に略語と元の語は 1 対 1 に対応するものだけではなく、ある語に対応する略語が複数ある可能性を認識する必要がある。

2.2 略語の作成法

略語の作成に関しては、大きく 2 つの処理に分けることができる。つまり多くの語に当てはまる規則的な表現が可能な処理と、その他の語に当てはまる個々に特有の処理である。

いずれの略語も「語の一部を省略して簡略にした語」という、略語の規則からは外れない。その相違は、統一的な処理は 1 つの規則で多くの略語の特徴を表現できるが、不規則的な略語はそれぞれに特有な方法により作成されるという点である。

2.2.1 略語作成における規則

多くの略語の作成の際に当てはまるものとして次の 4 つの規則が挙げられる。

- 略語は元の語から数文字を選び作成される
  言い換えると、略語に含まれている全ての文字は元の語に含まれているということである。

- 利用される語の順序は変わらない
  略語作成の際にその文字の前後関係は入れ替わらない。そのため、「独禁法」の例では文字の順を入れ替えた形の「禁独法」という略語

にはならない。

- 複合語は単名詞ごとの利用が多い
  複合語の場合は、略語に利用される文字として単名詞に区切り、それぞれの中から文字(情報)を選び、平均的に情報を含ませて作成することが多い。たとえば、「独占・禁止・法」の場合それぞれから1字ずつ利用され「独禁法」となる。

- 接頭辞の利用、右側を切り落とした部分の利用が多い
  単名詞の中でもその頭文字の利用が多い。また、各1文字のみではなく複数文字が取り出されることも多い。そのため、より一般的には「語の左側（始まりの文字）部分ほど利用されやすい」という表現ができる。たとえば「留守番・電話」の場合には「留電」ではなく「留守電」が利用されるパターンである。

もちろん、全てがこの規則に当てはまることはないが、非常に多くの略語がこの規則に当てはまる。そして、復元についてもこの作成規則を元に考えることができる。

2.2.2 不規則的に生成される略語

例外的に不規則に作成される略語も存在する。これらの略語は規則的でないため、個々に特有の作成法が存在することになる。そのため、全てを挙げて説明することは不可能であるので、いくつかの例を挙げ、その特徴を見る。

- 同じ語が異なる扱いを受ける場合
  たとえば、「電話」という語が含まれている語として「留守番電話」「公衆電話」「携帯電話」という3語を考える。この中で、「留守番電話」に関しては規則的に処理することができる。つまり、「留守番」「電話」と分け「留守電」という略語になる。しかし、通常「携帯電話」の略語としては「携帯」が用いられ、「公衆電話」については一般的には略語を用いない。つまり、同じ「電話」という語が含まれていてもそれぞれ異なる省略法が用いられていることになる。

- 俗称、段階を有する略語について
  また、俗称として元の語には含まれていない文字が付加される場合(「朝鮮民主主義人民共和国」に対する「北朝鮮」など)や、省略の方法が複数ある場合(「日米安全保障条約」に対する「日米安保条約」「安保条約」)など、規則的な処理と例外的な処理を併用するものもある。

これは、略語が日常的に人に利用されるものであるため、語呂など人の好みに依存する面を含むことを表わしている。

2.3 略語の復元

略語の復元とは、入力として与えられた略語に対し、その元の語を結果として出力するという作業をいう。そして、その方法としては、文書などから候補を見つけ、それらを絞りこむという作業になる。そのために、これまでに述べた略語の生成規則を踏まえ、また生成の際には見られないが、復元の際に見られる、復元独特の規則を導入し、略語の復元を考える。

略語は作成する際に規則を適用できるものと、規則が適用できない独特な方法によるものが存在した。しかし、現実的に復元を行うに当たっては、規則が適用できないものの復元は容易ではない。そこで、ここでは統一的に扱える規則を中心に復元法を見る。

略語の定義が「語形の一部を省略して簡略にした語」ということからも分かるように、略語に含まれる文字は元の語にも必ず含まれる。このような字面のレベルから次のような復元の規則を考えることができる。

- 語順に関する規則
  略語の作成の際に語の順序が入れ替わらないという規則があった。この規則はそのまま復

元の規則にも当てはまる。略語中の語は元の語の中でも同じ順で使われているということである。

- 複合語に関する規則
  複合語に関する作成の規則として、単名詞に分け、その頭文字、あるいは左側部分を利用することが多いという規則があった。この規則は復元においては、元の語を単名詞に区切り、それぞれの頭文字、あるいは左側部分とのマッチングを行うことで実現できる。

- 字種に関する規則
  複合語を分割して考えることに似ているが、通常カタカナと漢字が混ざる構成の語はカタカナと漢字を分けて考えることができる。例としては、「ソビエト連邦」は「ソビエト」から「ソ」を、「連邦」から「連」を取りだし、「ソ連」とする。

以上に挙げた字面からの方法が、一般的な略語復元の規則として考えられる。

3 日本語略語復元システム

3.1 システム概要

本研究での略語復元は、次の 3 つの作業から構成される。

- 新聞コーパスの利用
- 辞書の利用
- 規則の適用

この中で、まずはじめに最も簡易な規則を適用し、新聞コーパスおよび辞書から候補を広く集める作業を行う。そして、この方法により得られた候補に対し規則を適用し絞り込みを行う (図 1)。

3.2 新聞コーパスの利用

文書には多くの単語が含まれており、その中には略語の元になる語も含まれている可能性がある。

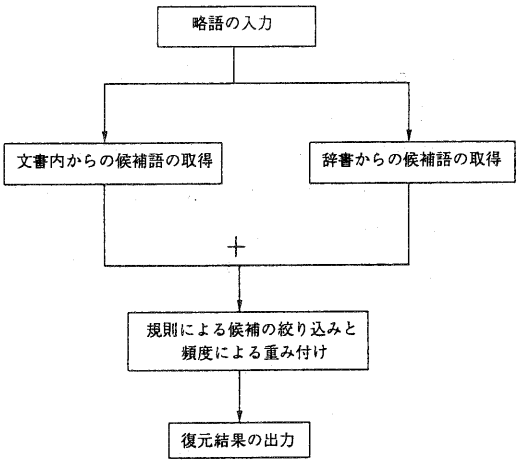

図 1: 略語復元処理の流れ

実際に、新聞を読んでいると略語に括弧書きでその元の語が含まれていることもある。そこで、本システムでは略語の元になる語の候補を、文書から探すという作業を取り入れる。

本システムで使用した文書は、1995 年の毎日新聞 [7] の 1 月から 6 月までのすべての記事で、その記事を茶筅 [6] を用いて形態素解析を行い、単語の区切りを与えておく。そして、略語内に含まれる文字を全て含む語をその元の語の候補として取り出す。この検索作業には、suffix array [2] [3] の方法を用いた。

このような作業により、文書から略語の元の語の候補を広く集める。

3.3 辞書の利用

文書から得られる候補には使用する文書の内容によって出現する語に偏りがあることや、同じ語が重複するという欠点がある。そこでこれを補うために、単語情報を合わせて利用する。

単語をコーパスとして利用する際に、最も簡単に利用可能なものとして、辞書がある。辞書は話題に左右されず、常に広い範囲の語を網羅するという長所がある。本システムでは、JUMAN[5] と茶筅 [6] で使用している名詞辞書と、広辞苑 [8] の

見出し語を合わせたものを辞書とした。その辞書から、略語と頭文字が同じであり、かつ略語内の文字が同じ順で含まれている語を候補として取り上げた。

3.4 候補の絞り込み

単純に、同じ語を同じ順で含んでいるという情報のみで復元が可能であるならば、それを結果として与えることができるが、実際にはこの条件を満たす語は多数存在することが多い。そこで、これら候補を絞り込み、その中でもより確からしさの高い語を選び出す必要がある。そのために次の規則を用いた絞り込みを行う。

1. 略語との完全一致語の除去
   略語と元の語が同じであることはない。そこで、完全な一致をする語、つまり略語自身を省くことで候補を絞り込むことが可能である。
2. 字種による絞り込み
   略語がカタカナのみから構成されている場合、その元の語もカタカナのみからできていることが多い。今回利用した略語延べ423語では、98.1 % がこの規則にあてはまる (表1)。つまり、異なる字種により構成されている語を省くことができる。

表 1: 略語の字種特性

	略語と元の語の字種が同数	字種が増える語数	字種が減る語数
割合	98.1%	1.9%	0%

(注) 答えが複数あるものはそれぞれ分けて数えている略語 423 語を対象とした

3. 文字数による絞込み
   略語は人が理解できる範囲で作り出される。過度に省略を行い曖昧性が出てしまっては略語としての機能がなくなってしまう。そのため、省略された 1 文字が持ち得る情報は長い場合でも 4 文字程度である。つまり、以下の式により候補を絞ることができる。

$$(\text{元の語の文字数}) \leq 4 \times (\text{略語の文字数})$$

この規則により文字数が必要以上に多い候補を外すことが可能となる。今回利用した略語では、88.9 % までがこの規則を満たす (表2)。

表 2: 略語の元の語に対する単語長

元の語の字数 / 略語の字数	1 ～ 2	2 ～ 3	3 ～ 4	4 ～ 5	5 ～
割合 (%)	16.1	44.2	28.6	8.5	2.6

(注) 元の語が複数あるものはそれぞれ分けて数えている

4. 元の語内でどのように含まれているか
   略語に含まれる文字が、元の語の中にどのように含まれているかを考えることで候補を絞ることができる。たとえば、「京大」という略語があるとき、元の語の中で「東京大学」のように連続的に利用されているよりは、「京都大学」の様に不連続に利用されている可能性が高い。つまり、連続的に利用されている語を省くことができる。今回の略語では、83.0 %がこの規則を満たす (表3)。しかし、これには例外が多いため、他の規則だけでは候補が絞り切れない時のみ利用するなど、慎重に扱う必要がある。この規則の例外としては、「最高裁判所」の「最高裁」などが挙げられる。

表 3: 略語字の含まれ方

	全略語対象	カタカナのみの略語を除く
全略語数	423 語 (注)	402 語
略語内の文字全て連続的に使用	17.0%	14.2%
不連続で使用	83.0%	85.8%

(注) データは今回利用した略語 (元の語が複数あるものは分けて数えている)

3.5 候補の重み付け

本研究では、候補を複数与えるため、それらに対し優先順位を付け、候補の中でどの語がより適切であるかを示す。本来、重み付けを行う際に、最も適切と考えられる方法として、文書内での利用法を調べる方法がある。つまり、文脈情報の利用である。略語と元の語は意味的に同じものであるため、同じ形で利用されている語をより可能性の高いものとして扱うということである。

しかし、本研究においては入力を単語のみの情報として与えるため、文脈情報が利用できない。そこで、今回は出現頻度を利用してその重み付けを行った。これは、例外的な語が低頻度で出現するため、このような例外を省くことを考えている。より一般的な語を可能性の高い候補と見ることになる。

4 復元実験

実験は、404 語の入力略語を対象に本システムを適用し復元を試みた。また、規則を全て、あるいは、一部を適用し以下のような 4 通りの実験を行った。

実験 1

全ての規則を適用して略語の復元を行う。全ての規則とは、以下の 4 つである。

- 元の語が略語内の文字を全て、同じ順で含むこと
- 略語と元の語を構成する字種が等しいこと
- 元の語の文字数が略語を構成する字数の 4 倍以内であること
- 略語内の文字が元の語の中で不連続的に含まれていること

実験 2

全 4 規則のうち、次の 3 規則を用いて復元を行う。

- 略語内の文字を全て、同じ順で含むこと
- 略語と元の語の字種が等しいこと
- 字数の制限が 4 倍以内であること

実験 3

全 4 規則のうち、次の 3 規則を用いて復元を行う。

- 略語内の文字を全て、同じ順で含むこと
- 略語と元の語の字種が等しいこと
- 不連続的に含まれていること

実験 4

全 4 規則のうち、次の 3 規則を用いて復元を行う。

- 略語内の文字を全て、同じ順で含むこと
- 字数の制限が 4 倍以内であること
- 不連続的に含まれていること

5 結果と考察

実験結果を表 4 に示す。ここで完全一致とは、元の語がきちんと復元されたものが候補に入っている場合で、「参院」「パソコン」「独禁法」「安保条約」「理研」などである。拡張一致とは、候補語には正解が入っていなかったものの、候補語から容易に元の語が推測できる場合で、「国土法」(候補では「国土利用計画法違反」が出力された)「科捜研」(候補では「同庁科学捜査研究所」)などがある。これは、新聞文書中の名詞の連鎖を複合名詞として利用した際に、取り出されたものである。

5.1 復元の精度 (再現率) について

再現率 (表 4 での完全一致および完全+拡張一致) が示す通り、入力略語に対して 7 割は元の語が含まれている結果を与えていることになり成功といえる。一方、全実験において 25%ほどの略語で復

表 4: 略語復元実験の結果

	実験 1	実験 2	実験 3	実験 4
入力数	404 語	404 語	404 語	404 語
完全一致	71.8%	73.5%	74.8%	71.0%
拡張一致	4.2%	4.2%	3.7%	4.5%
完全＋拡張	76.0%	77.7%	78.5%	75.5%
適合率	29.0%	19.0%	25.0%	25.9%
拡張的適合率	78.6%	79.2%	80.5%	78.4%

元に失敗している。この原因には以下の点が挙げられる。

- コーパス内に元の語が存在しない場合
  これは新聞コーパスおよび辞書に元の語が含まれていない場合であり、これらへの対応として略語に合うコーパスを選択することや、略語単位ではなく文字単位での復元を行った。しかし、これらの対応ではごく少数の語に対しては効果が見られたが、全体としての再現率を大きく上げるほどの効果は見られなかった。

- 規則適用により候補を外れる略語
  コーパス内に元の語はあったが、規則を適用する中で候補から外れて出力されなかったという語もある。これらの原因としては、不規則的に生み出された略語であることが考えられる。つまり、全ての略語にあてはまる規則ではなく多くの略語にあてはまる規則を用いたため、その例外が復元できなかったということである。しかし、今回の規則が不適切な語を多く排除しているため、これらの規則利用は妥当であったといえる。

5.2 適合率について

結果を複数与えることを許したため、その適合率が低くなったことも事実である。複数の結果を認めるということは、例えば各略語に対し 5 語の結果が与えられたと仮定した時、正解が各入力の結果に 1 語含まれている場合でも適合率は 20%ということになる。出力が 5 語であり、その中に 1 語、正解が含まれているならば、実用上は問題がないと考えられる。しかし、実用上問題がないと考えられるこのような場合においてもその適合率は 20%となるため、今回の適合率も成功と言える数値である。

拡張的適合率は

$$\text{拡張的適合率} = \frac{\text{正解を含む略語数}}{\text{出力が 1 つ以上あった略語数}}$$

というもので、正解をどれほど含んでいるかを略語ごとに捉えた数値である。結果が 1 語以上与えられた略語については、8 割程の精度で正しい元の語を含んでいた。

ここで、平均出力数を見ると表 5 のようになっている。実際に、4 から 6 語程度となっていることが分かる。また、平均出力数は全ての規則を適用

表 5: 候補となった語の平均出力数

	実験 1	実験 2	実験 3	実験 4
平均出力数	3.7 語	5.7 語	4.3 語	4.0 語

した場合、つまり実験 1 のときに最も低い値を示した。このとき、再現率は大きく下がらないため、全ての規則が有効に働いているといえる。

5.3 重み付けについて

重み付けの精度をみるため、候補が 2 語以上出力されたもののうち、最大重みが正解語であった割合を表 6 に示す。この結果から、重み付けにより 7

表 6: 重み最大が正解語であった割合

	実験 1	実験 2	実験 3	実験 4
割合	72.2%	69.2%	73.0%	71.8%

割程正しい元の語が最優先されていることが分かる。これにより、頻度による重み付けが正しい結果を与えていたと言える。しかし、「ダイヤ」など略語のみからは可能性が複数存在する (「ダイヤモン

ド」「ダイヤグラム」)語があるため、略語の復元においては重み付けは最も適切な語を選ぶためではなく、不適切な語を削除するために利用することが妥当であるといえる。

また、3割に対しては正確な結果を与えていなかったが、この原因として次のことが挙げられる。

1: 新聞に登場しやすい語の重みが高くなる(「外為法違反」など)

2: 低頻度の語ばかりのときは重みに差が出ない

3: 対象が多いときは、全てが低頻度になる

重みの数値の利用法には一般に2通りが考えられる。1つは順位付けにより特定の順位までを候補として抜き出す方法であり、もう1つはある閾値以上を候補として扱う方法である。今回は後者を利用した。この方法では、前者とは異なり、出力語数にばらつきが見られるが、余分な語を省き高頻度の候補のみを確実に出力していたため成功といえる。

ただし重み付けには確実といえる方法はないため、最も確からしい1語を選び出すためではなく、不適切な語を排除するために利用するのがよいだろう。

6 結論及び今後の課題

今回日本語の略語復元を行ったが、新聞コーパスと辞書を用い、また簡単な規則を適用させることである程度略語の復元が可能であることがわかった。

今回の方法では、その出力候補数が平均3から4語程度となった。この出力数を1語に限定するという作業は非常に困難である。この程度の出力数は、十分実用に耐えうると考えられるため、結果に対する最終判断は、本システムにより選び出された限られた候補から利用者が選ぶことで解決する、という形での実現となった。

今後の課題としては次のような点が挙げられる。

- 文脈情報の利用
  今回の方法では、孤立略語からの復元であった。しかしながら、この方法では限界があることは明らかである。そこで、その略語がどういう文脈で使用されたかといった情報を利用することでその精度を上げることができると考えられる。

- 略語の判定
  入力は必ず略語が与えられるという仮定のもとに本システムは作られている。しかし、実際は入力語が必ずしも略語とは限らない。そこで、入力の際にその語が略語であるかの判定を行う必要が生じる。これにより文書からの略語の抽出、復元などの発展が考えられる。

- 英語への対応
  今回は、原則として日本語を対象に考えている。英語にも数多くの略語が存在するので、日本語だけでなく英語にも対応することを目指したい。

参考文献

[1] Neil C. Rowe, Kari Laitinen, "Semiautomatic disabbreviation of technical text", Information Processing and Management, 31, no.6, pp.851–857, 1995.

[2] U. Manber, G. Myers, "Suffix arrays: A new method for on-line string searches", SIAM Journal on Computing, Vol. 22, pp.935–948, 1993.

[3] sufary ホームページ
http://cactus.aist-nara.ac.jp/lab/nlt/ss/index.html

[4] Gregg M. Stum, Patrick W. Demasco, Kathleen F. McCoy, "Automatic Abbreviation Generation", Proceedings of the Fourteenth Annual RESNA Conference, pp.97-99, 1991.

[5] JUMAN ホームページ
http://pine.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[6] 茶筅ホームページ
http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html

[7] CD-毎日新聞'95, 日外アソシエーツ.

[8] 広辞苑 第4版 (電子ブック版), 岩波書店.