日本語辞書整備のための日本語カタカナ複合名詞の自動分割

中澤 敏明　　河原 大輔　　黒橋 禎夫
東京大学工学部　東京大学大学院情報理工学系研究科
{nakazawa, kawahara, kuro}@kc.t.u-tokyo.ac.jp

1　はじめに

現在の日本語形態素解析では、人手により整備された辞書を用いている [2]。漢字や平仮名については、辞書が十分に整備されており、解析精度は非常に高いものとなっているが、問題となるのはカタカナ語である。カタカナ語は生産性が非常に高く、専門語、新語などが多数あり、人手による逐次的な辞書整備は困難である。

これまでの日本語形態素解析では、辞書登録されていないカタカナ列で、辞書にあるカタカナ語の組み合わせとして解釈できないものは、全体を一語の名詞と推測していた。また、単語分割では基本的に長い語が優先される。これらの結果、たとえば「トマトソース」というカタカナ列は、「トマト」と「ソース」が辞書登録されており、「トマトソース」が登録されていないという場合にかぎって、「トマト ＋ ソース」と正しく分割されることになる。「ソース」のみが登録されている場合 (辞書が不完全な場合) や、「トマトソース」まで登録されている場合 (辞書が冗長な場合) には「トマトソース」が一語となってしまう。

しかし、このような扱いは明らかに不十分である。つまり、「トマトソース」を一語としてしまうと、テキストを深く解析するために必要となる、「トマトソース」が「ソース」の一種であるという情報が利用できない。また、「トマトソース」を含むテキストを「トマト」や「ソース」などの語で検索することもできないことになる。かといって、部分文字列マッチングのような粗い方法を使ってしまうと、「スライス」に「ライス」「イス」などがマッチすることになり、検索の精度を悪化させてしまう。

このような問題を解決するために、本論文では、カタカナ語集合に対して、各カタカナ語が複合語か否かを自動認識し、独立の語だけからなる正確で簡潔な辞書を自動構築する方法を提案する。本手法は、ある程度の規模のコーパスから単純にカタカナ列を取り出し、その頻度を計数したデータを入力とする。以下ではこのデータを基礎データとよぶ。

ラーメン	28727	メニュー	14766
スープ	20808	エスニック	14190
レシピ	16436	サラダ	13632
カレー	15151	...	

このような基礎データと、一般的に利用できる和英辞書と英語コーパスを用いて、各カタカナ語が複合語であるかどうかを判断する。

我々の提案手法は、次の 3 つの手法を組み合わせるものである[1]。

- 和英辞書を用いる方法
- 英語コーパスと和英辞書を用いる方法
- 基礎データ内の関係を用いる方法

まず、カタカナ語の大半が英語に由来することから、英語の情報をできるだけ利用することを考え、和英辞書と英語コーパスを用いる。ただし、それだけでは十分に再現率が高くならないことから、基礎データ内で分割されるものは複合語と見なすことを、閾値をもうけて行う。上記の手法は、適合率の高いものから再現率の高いものという順になっており、これらを適切に組み合わせることで、全体として、適合率・再現率ともに非常に高い処理を実現した。

2　和英辞書を用いる方法

まず、和英辞書の持つ情報を最大限に活用し、あるカタカナ語については分割することを、あるカタカナ語については一語であることを確定する。ここで処理された語は、あとの手法では処理されない。

基本的なアイデアは次のとおりである。例えば、「トマトソース」に関して次のような情報が辞書にあるとする。

トマトソース = tomato sauce

[1]なお、動植物名、食品名などの、日本固有の単語でカタカナ表記されるものについては、日本語の国語辞典に載っているものは一語として扱う、という方法で対処した。

トマト = tomato　　ソース = sauce

このとき、訳語が複数の語からなり、その各訳語とカタカナ文字列の部分対応が正確にとれることにより、「トマトソース」は「トマト + ソース」という複合語であることがわかる。

逆に、例えば「サンドウィッチ」という語は、「サンドウィッチ = “sandwich”」と一語の訳が与えられているので、これは一語であることを確定する。

このように、辞書の情報を単純に利用することによって、複合語か否かがわかるが、実際にはもう少し細かな取り扱いをしている。例えば複数語訳で、「モルネソース」のように、部分の対応が完全にとれなくても、末尾語の対応がとれるものは複合語であるとする。

モルネソース = Mornay sauce
ソース = sauce

3 英語コーパスと和英辞書を用いる方法

辞書には基本的な複合語は登録されているが、コーパス中に出現する複合語はそれよりもはるかに多く、辞書を直接用いるだけでは不十分である。そこで、基本的な日英の対応関係は和英辞書を用い、それを基にして英語コーパスをチェックすることによって、複合語か否かを調べるという方法を考案した。

注目するカタカナ語に対して、和英辞書に登録されているカタカナ語の組み合わせへの分割を考え、それぞれ訳語に変換することにより、元のカタカナ語の訳語の候補を作る。これを、考えられるすべての分割と、すべての訳語変換について行う。そして、それらの英語コーパスでの頻度を調べ、その中の最大のものが閾値以上出現していれば、そのカタカナ語はそのように分割されると決定する。英語コーパスとしては Web を用い、サーチエンジンでのヒット数を頻度と考える。

例えば、「パセリソース」は次のような計 3 種類の候補について Web 頻度を調べる。

パセリ + ソース parsley source:554
パセリ + ソース parsley sauce:20600
パセ + リソース pase resource:3

この結果、上記のように、「パセリ + ソース（parsley sauce）」が高頻度に出現しているので、この分割が正しいものと判断する。またこの場合、パセリソースの適切な訳語も得られることになる。

本論文で問題としているのは、あるカタカナ語が複合語であるかどうかなので、分割するかどうかの閾値が重要である。Web にはゴミも多く、おかしな分割、訳語であっても出現があり、低頻度の分割は信頼できない。例えば次のような不適切な分割にも低頻度の出現がある。

デミ + グラス demi glass:207
バン + バンジー van bungee:159

そこで、長いカタカナ語は分割される可能性が高いことを考慮し、C/N^L という閾値を設定した。ここで L はカタカナ語の長さであり、C と N はそれぞれ適当な定数である。

4 基礎データ内の関係を用いる方法

和英辞書や英語コーパスに基づく分割は、信頼度の高い高精度な手法であるが、構成語が和英辞書に載っていて、かつその語が英語としても適当な複合語になっていなければ適用することができない。しかし、新語や専門用語であったり、カタカナ書きすることが一般的でない場合には和英辞書には載っていない。また日本語でのみ用いられる和製複合語というものも存在する。

そこで、基礎データの中で、適当に分割した部分カタカナ語がそれぞれにある程度以上の頻度で出現すれば、そのように分割されると推測する。

問題になるのは閾値である。コーパス中には相当種類のカタカナ語が出現するので、本来一語である単語でも、何らかの分割した解釈が存在してしまい、単純にはほとんどの語が複合語とみなされるからである。

そこで、基本的には元のカタカナ語の頻度 (F_o) と、分割構成語の頻度の相乗平均 (F_g) を比較し、後者の方が大きければ分割すると判断する。ただし、前節と同様に、長い単語は複合語である可能性が高いということを考慮して、長さに応じて閾値を調整する。すなわち、調整された相乗平均値を F'_g として、

$$F_o < F'_g, \qquad F'_g = F_g/(C/N^l + \alpha),$$

という条件が満たされれば分割すると考える。ここで l は分割構成語の語長の平均であり、C と N と α はそれぞれ適当な定数である。

複数の分割が考えられる場合は、その中で最も F'_g の値が大きい分割を考える。また、2 語への分割と 3 語への分割が考えられる場合は、少ない語数への分割、すなわち 2 語への分割だけを考える。これは、多くの短い語への分割では、短い語の出現頻度が極端に大きく、相乗平均への頻度比較が不適切になる場合があるためである。

以下にいくつかの例を示す。

イタリアンレストラン:$F_o = 207$
↔ イタリアン:1421 + レストラン:7922 ($F_g = 3355$)

イタリアン:$F_o = 421$
↔ イタ:91 + リアン:11 ($F_g = 31$)
↔ イタリ:7 + アン:301 ($F_g = 45$)

「イタリアンレストラン」は「イタリアン」という語が和英辞書にないため、前節の方法では分割されない。一方、「イタリアン」は「イタ + リアン」、「イタリ + アン」などの分割が考えられるが、いずれも F_g の値が小さく、このような分割は採用されない。

5 実験と考察

5.1 実験結果

実験は、新聞記事 12 年分（580 万文）中のカタカナ列で頻度 2 以上のもの、87,000 語と、料理について書かれた Web ページを集めたコーパス (280 万文) 中のカタカナ列で頻度 2 以上のもの、43,000 語の 2 セットについて行なった。

2 セットともに 500 語からなる評価セットを用意し、それぞれの語に正解分割位置を人手で与えた (複合語でなければ分割位置を持たない)。そして、これを自動分割の結果と比較し、再現率・適合率を求めた。評価は、単語単位ではなく、分割位置単位で行なった。なお、500 語について平均の分割数は新聞セットが 1.39、料理セットが 1.62 であった。

これまで述べたとおり、提案手法は次の 3 つの処理からなる。

1. 和英辞書を用いる方法 (D)
2. 英語コーパスと和英辞書を用いる方法 (C)
3. 基礎データ内の関係を用いる方法 (R)

各処理の有効性を調べるために、D のみ、D+C、D+R、D+C+R の 4 種類の処理を行い、比較した。なお、閾値は 2 の方法については $400,000/2^L$、3 の方法については $F'_g = F_g/(2,500/4^l + 0.7)$ とした。また、和英辞書としては、英辞郎 (見出し語数 93.1 万、内カタカナ語 13.7 万) と edict(見出し語数 14.0 万、内カタカナ語 1.4 万) を用いた。

この結果をまとめたものを表 1 に示す。この表に示すとおり D+C+R の方法によって、適合率、再現率ともに高い結果をえることができている。

また、図 1 は新聞データ、料理データそれぞれの頻度 10 以上のカタカナ語について、分割された語、分割されない語、そのうち形態素解析辞書にすでに登録されている語の数を、単語長ごとに示したものである。新聞データについては、14,000 語中 4,000 語が分割され、料理データについては、4,900 語のうち 2,500 語が分割された。

5.2 考察

表 1 に示した通り、辞書を用いる方法は適合率は高いがこれだけでは再現率が低い。英語コーパスを用いる方法と基礎データ内関係を用いることによって、適合率・再現率ともに非常に高い精度がえられている。

誤りの原因を分析すると次のようになる。

適合率の問題、すなわち誤った分割、過分割される場合は、まず和英辞書で一語であると認定できず、それが後続する 2 つの処理で分割される。辞書で一語と認定できないものは主に次のようなものである。

- **新語や、あまり日本語で用いられない語**

  セル + ライト cell light:15100 > 閾値:12500
  (正しくは “cellulite”)

  シュレッドチーズ:$F_o = 24$
  ↔ シュ:41+レッド:112+チーズ:7199 ($F'_g = 143$)

- **表記揺れの問題**

  代表的表記は辞書に載っており、一語であるなどのより正確な情報がえられるが、表記揺れにはそのような情報がなく、不適切に分割される。

  プラスティック:$F_o = 48$ (代表表記:プラスチック)
  ↔ プラ:67 + スティック:224 ($F'_g = 143$)

- **固有名詞の問題**

  固有名詞は和英辞書では十分にカバーされていない。NE 処理などと本手法を統合して再検討する必要がある。

  コネティカット:$F_o = 108$
  ↔ コネ:177+ティ:166+カット:4144 ($F'_g = 108$)

一方、再現率の問題、すなわち本来分割すべきものが分割されないものについては、原因は次のようなものであった。

- 特に短めの単語については、複合語なのか否かの基準の問題があり、日本語では複合語と判断しても、和英辞書で一語となっていることが原因となっていた。
  フレックスタイム flextime
  プールサイド poolside

- 適合率の場合と同じく、分割されるべき語が辞書にないと、Web を引くことができず、さらに、データ内関係でも切られないということがあった。

  ベイエリア:$F_o = 163$ (ベイが辞書未登録)
  ↔ ベイ:116 + エリア:1377 ($F'_g = 127$)

- また、もちろん、辞書にあってもすべてが閾値を越えない場合もある。

表 1: 実験結果 (適合率/再現率)

	D	D+C	D+R	D+C+R
新聞データ	1.0/0.822	0.996/0.909	0.986/0.945	0.985/0.949
料理データ	1.0/0.717	1.0/0.836	0.990/0.948	0.991/0.956

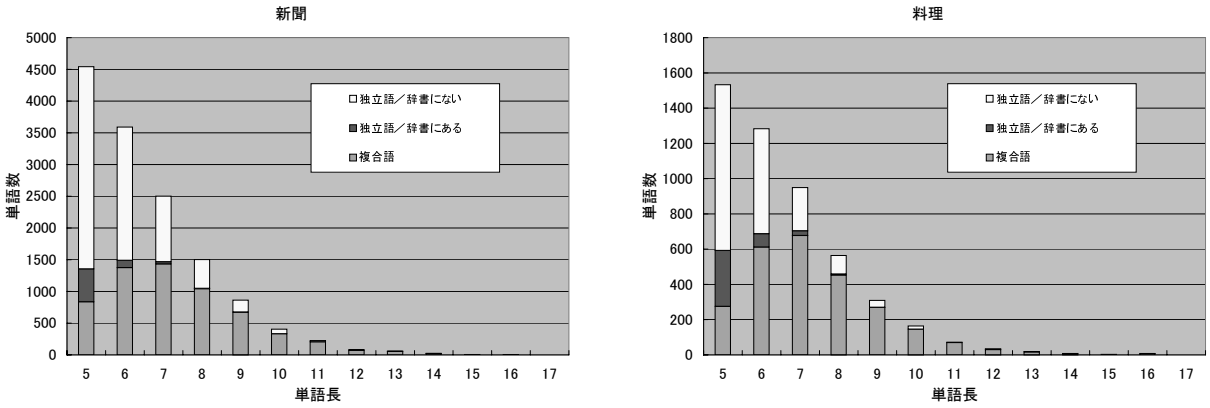

図 1: 複合語と独立語の文字長による分布

ペパー + ミント pepper mint:5400 < 閾値:6250
ペパーミント:$F_o = 41$
↔ ペパー:8 + ミント:56 ($F'_g = 16$)

6 関連研究

我々の知る限り、大規模に表音文字複合語の自動分割を扱った研究はこれまでのところ行なわれていない。本研究で扱ったものと同じ問題はドイツ語の複合語でも発生し、この手法を適用することが考えられる。

一方、我々の手法を発展させるために利用できる関連研究を以下に挙げる。まず、和英辞書の利用において、その中の英語と日本語が transliteration の関係 (英語の発音をカタカナで表わしたもの) になっているかどうかは重要であり、これには Knight らの研究が利用できる [1]。

また、カタカナ語の表記揺れも重要な問題である。これには多くの研究があり、最近のものとしては増山らの研究がある [3]。

我々の手法の英語コーパスに基づく方法によって獲得される情報は、訳語学習と考えることもできる。英語コーパスを用いた訳語学習としては宇津呂らの手法などがあり、今後、比較検討を行なう予定である [4]。

7 おわりに

はじめに述べたように、ある規模のコーパスさえ与えられれば、我々の提案した手法を用いてカタカナ複合名詞を自動分割することにより、正確で簡潔なカタカナ語辞書を構築することができる。カタカナは英語の表音文字として使われることが多いため、我々は和英辞書と英語コーパスを利用した。また、訳語を元にした手法と、頻度の関係を元にした手法を用いることにより、適合率・再現率ともに高い結果が得られた。

これらの結果は、すでに日本語形態素解析の辞書の拡張に利用している。今後は表記揺れの吸収、及び、固有名詞判定手法を統合する予定である。

参考文献

[1] Kevin Knight and Jonathan Graehl. Machine transliteration. *Computational Linguistics*, 24(4):599–612, 1998.

[2] Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of The International Workshop on Sharable Natural Language*, pages 22–28, 1994.

[3] Takeshi Masuyama, Satoshi Sekine, and Hiroshi Nakagawa. Automatic construction of Japanese katakana variant list form large corpus. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 1214–1219, 2004.

[4] Takehito Utsuro, Kohei Hino, Mitsuhiro Kida, Seiichi Nakagawa, and Satoshi Sato. Integrating cross-lingually relevant news articles and monolingual web documents in bilingual lexicon acquisition. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 1036–1042, 2004.