連語辞書の自動作成と評価

5F−3

山田洋志　大竹暁子
(NEC C&Cシステム研究所)

1 はじめに

日本語解析に連語(単語の共起関係)を利用することで解析精度を向上させることができる。しかし、実際に連語を利用するにあたっては、連語のデータをどのようにして収集するかという点が大きな問題となる。

田中は、技術文献・新聞記事などから手作業によって共起関係データを抽出している[1]。手作業による作成はデータの信頼性が高いが、作成のためのコストが非常に大きいという問題がある。それを解決するために、テキストを解析し自動的に共起関係を抽出する方法も提案されている。この方法の問題は、解析の精度などの関係から誤ったデータが抽出されることにある。そこで誤りを減らすために、抽出対象として解析結果の曖昧さの少ない部分を使ったり[2]、対象テキストを限定する[3]などの対策がとられている。

筆者らは、かな漢字変換の解析精度を向上させるために大量の連語を用意するという方針を立てて検討を行っており[5]、そのための手段のひとつとして形態素解析を利用した自動抽出方式を検討してきた[4]。本方式では誤ったデータの抽出を抑さえるために、単語の品詞に基づいたヒューリスティックな規則を利用している。

本稿では、ヒューリスティックな規則を用いた抽出方式を提案し、作成した辞書の評価結果について報告する。

2 連語辞書の作成方式

テキストを形態素解析し、解析結果から抽出条件に合う単語の並びを抽出して連語辞書を作成する。

連語抽出の条件は、単語列の品詞の組み合わせに基づいている。抽出条件の決定には以前に行った実験結果[4]を参考にして、以下の点を考慮した。

- 誤った連語の抽出はできるだけ減らす。
- こと・ものなど、連語としての効果が小さいものは抽出しない。そのために抽出対象を自立語とした(「こと・もの」は形式名詞という品詞区分にしている)。
- 漢字1文字の単語を含んだ連語は、解析誤りの原因となりやすいため抽出対象から外した。

抽出条件の詳細を2.1節、2.2節に記す。

2.1 文節内の条件

自立語(名詞、固有名詞、サ変名詞、動詞、形容詞、形容動詞、副詞)を含む文節を、連語データ作成の対象とする。さらに、各文節の自立語に付く付属語、活用形も考慮する。

連語の前部分となる文節についての条件を挙げる。

1. 接辞が付かない。
2. 名詞・固有名詞……単独または「特定の助詞」とともに文節を構成する(「特定の助詞」={を、の、が、に、で、と、や}。係助詞(に、も)が付いても可)。1文字の名詞・固有名詞は除く。
3. サ変名詞……2に同じ。動詞として使われている場合は文節末尾が連体形であること。
4. 動詞……文節末尾が連体形であること。
5. 形容詞……文節末尾が連体形か、連用形であること。
6. 形容動詞……5に同じ。名詞的用法の形容動詞は除く。
7. 副詞……1文字の副詞は除く。

連語の後部分となる自立語は、接辞が付かないことが条件である。

2.2 2文節間の条件

隣り合う2文節が前節の条件に合っていたなら、表1の規則にしたがって連語データ抽出の対象とするかどうかを決める。

表 1: 連語抽出条件

前の文節＼後の文節		名詞	サ変名詞 名詞	サ変名詞 動詞	形容詞	形容動詞	動詞	副詞
名詞・サ変	の,と	末	末	○	※	※	○	×
	を,が,に,で	×	×	○	※	※	○	×
	や(と)	末	末	×	×	×	×	×
	単独	○	○	○	○	○	○	×
	連体形	○	○	○	○	○	○	×
動詞(連体形)		○	○	×	×	×	×	×
形容詞(連体)		○	○	×	×	×	×	×
形容詞(連用)		×	×	○	○	○	○	×
形容動詞(連体)		○	○	×	×	×	×	×
形容動詞(連用)		×	×	○	○	○	○	×
副詞		×	×	○	○	○	○	○

○……後方の文節のうち一番近い単語と連語とする。
×……連語としない。
末……後方が名詞・サ変名詞からなる複合語の場合、末尾の単語を連語とする。
※……後方の単語が連用形以外のとき、連語とする。

ただし、前の文節が句読点で終わっている場合、“名詞＋固有名詞”、“人名＋名詞”の組み合わせになっている場合は抽出対象としない。連続した文節から抽出するのを原則としたが、“名詞＋格助詞”と動詞の組み合わせについては、間に他の文節があっても抽出した。

3 連語辞書作成実験

2章に述べた方法で、連語辞書自動作成の実験を行った。

14,738,975文字のテキストを形態素解析し、363,300レコードからなる連語辞書を作成した。図1に連語辞書の一部を示す(見やすいように実物とは形式を変えてある)。

An Automatic Collocation Acqisition Method and its Estimation

Hiroshi YAMADA and Akiko OHTAKE
C&C Systems Research Laboratories, NEC Corporation

表記 (品詞) −関係− 表記	(品詞)	頻度
衛星 (名詞) − が − カバー	(サ変)	01
衛星 (名詞) − − 株式会社	(名詞)	01
衛星 (名詞) − に − 関	(動詞)	01
衛生 (名詞) − − 環境	(名詞)	01
衛星 (名詞) − の − 管制センター	(名詞)	02
衛生 (名詞) − − 鑑定	(サ変)	02

図 1: 自動作成した連語辞書の一部

連語辞書の品質を調べるため、連語辞書から無作為に726レコードを抽出し、連語として適切かどうか調査した。結果を表2に示す(前回の実験時の割合も併記した)。

表 2: 連語の分類

【適切】 例：愛着/の/深い、跡地/に/完成	70.5% 前回 50.5%
【やや不適切】 例：赤字/が/のぼる、上げ/を/つくる	17.6% 前回 15.9%
【不適切】 例：位置/ /にあう、打ち/ /返す	11.8% 前回 33.5%

4 かな漢字変換への適用評価

筆者らが開発したかな漢字変換システムでは、共起関係にある2単語を登録した辞書(連語辞書)を使用し、隣接する2単語で利用している[4]。3章の連語辞書をかな漢字変換で試用して効果を調べた。結果は表3の通りである。変換率は長文節単位に計算し表記の揺れは正しい変換とした。

表 3: かな漢字変換の文節単位変換率

分野	文節数	連語未使用	連語使用
教科書	12,334	85.7%	86.5%
新聞記事	13,773	83.2%	84.0%
日常文書	8,454	83.9%	84.3%
業務文書	8,196	87.0%	87.4%
専門分野	10,491	87.3%	87.7%
合計	53,248	85.3%	85.9%

全体として0.6%正解率が向上している。

4.1 連語辞書の効果

前節の結果の一部について、連語辞書を使用したことによって変化のあった箇所を調べた。誤りから正解に変わった箇所が126、正解から誤りに変わった箇所が32である。その他、揺れの範囲内で表記が変化した箇所(変換率には変化なし)が214であった。

4.2 他の辞書との比較

自動作成した連語辞書を人間が直接作成した連語辞書と比較する。

辞書1は、人間が収集したデータ[1]から作成した連語辞書、辞書2は、自動作成した連語辞書である。

表4は、この2種類の連語辞書を用いて同一テキストのかな漢字変換を行った場合の効果の比較である。表4で、"連語数"は連語辞書の見出し数、"正変換箇所"は連語辞書を使用することで正しく変換された箇所の増加数である。

表 4: 連語の効果の比較

辞書	作成方法	連語数	正変換箇所
1	手作業	275,291	1,270
2	自動	363,300	320

自動作成した辞書では、連語数が多いにもかかわらず、正変換箇所の増加が少ない。

5 考察

3章で述べたとおり、自動作成した連語のうち、連語として適切なものの割合は70.5%であった。改良前の方式で作成した連語辞書では50.5%であり、抽出条件の改良は不適切なデータの削減に効果があった。

不適切だと判断した連語のうち、数の多いものを挙げる。

"やや不適切"のうちでは、隣り合っていない2文節から抽出されたものが目立つ(例:「赤字が一億円にのぼる」から、「赤字/が/のぼる」を抽出)。この条件の見直しが必要である。

"不適切"のうちでは、形態素解析段階での失敗が原因のものが多い(例:打ち/返す、置き/どころ)。改善のため、現在、形態素解析自体の改良を行っている。

かな漢字変換での実験で正解から誤りに変わった箇所に、連語で自立語を分割して誤りになったものがある(例：正/ホルモン/の/働きによって、他/維持/の/話し)。これについては、かな漢字変換側との調整が必要である。

4.2節の結果は、今回自動作成した辞書の品質が人間が直接作成した辞書の品質には及ばないことを示している。したがって、現状では大量のデータを用意して品質をカバーするか、あるいは、直接使用せずに人間による作成の補助として使用するのが有効であろう。

6 まとめ

形態素解析を用いてテキストから自動的に連語を抽出する方式の提案・評価を行った。この方式では、単語の品詞に基づく複数のヒューリスティックな規則の組み合わせで、誤ったデータの抽出を押えている。今回、約1400万文字のテキストから約36万レコードの連語辞書を自動作成したところ、適切な連語の割合が70.5%となり、以前の実験より20%改善された。

また、かな漢字変換で使用した場合の効果を調査した。人間が直接作成した辞書の効率には及ばないものの、正変換率を0.6%向上させることができた。

現在は、誤ったデータ抽出を減らすための大きな要因である形態素解析の精度向上のための改良を行っており、その後抽出規則の見直しを行う予定である。また、作成した辞書の利用方法についても検討する。

参考文献

[1] 田中他,情処40全大5F−1,1990
[2] 本間他,情処誌vol.27,No.11,1986
[3] 中島他,情処38全大2E−6,1989
[4] 山田他,情処42全大5Q−3,1991
[5] 山田,自然言語処理研究会 87-5,1992