弱抑制による連鎖共起表現の抽出とそれに基づく離散共起表現の抽出

内野　一　　池原　悟　　白井　諭
ＮＴＴコミュニケーション科学研究所

１．はじめに

機械翻訳などの自然言語処理において、固定的な言い回しの表現を収集するため、大規模な言語データベースから共起表現を収集する方法が期待されている。共起表現には、いくつかの単語が連続したタイプ（連鎖共起表現）と、２種類以上の表現が文中の離れた位置に共起するタイプ（離散共起表現）があるが、単語単位の処理だけでは正しく扱うことが困難で一定の構造を持つ文型に着目する必要がある処理においては、後者のタイプを自動的に収集することが特に有効と考えられる。n-gram統計処理を用いて前者の表現を抽出する方法が提案されている[1]が、断片的な文字列がかなりの割合で混在し総頻度が大きくなりすぎるという問題がある。離散共起表現を抽出するには、基となる連鎖共起表現を組み合わせ計算が可能な量に抑えることが必要となるが、抑制を強くすると今度は最終的な離散共起表現の抽出量が少なくなってしまうという問題がある。本稿では、連鎖共起表現の抽出に弱めの抑制をかけることによって、この問題を解決する方法を提案する。

２．連鎖共起表現の抽出

２．１　従来の技術

n-gram統計処理で得られる文字列から、長単位の表現に含まれる部分を除去する事によって、連鎖共起表現の総頻度を抑え、離散共起表現の抽出を可能にする手法が提案されている[2]。しかしながら、この方法を用いて連鎖共起表現を抽出した場合、離散共起表現の抽出結果には要素数３以上で、度数３以上の組がほとんど現れない。また、基本的に表現を長い単位で抽出するため、基本的な短い語で構成される呼応表現が抽出され難いという問題がある。

例えば、「まだ…走らない」「まだ…動かない」という表現ががそれぞれ２回、「まだ…ない」という表現が１回含まれているデーターベースから離散共起表現を抽出すると［まだ　走らない］［まだ　動かない］の表現が頻度２で抽出される。しかしながら［まだ　ない］という呼応表現は実際には頻度５であるにも関わらず抽出されない。

２．２　弱抑制型連鎖共起表現抽出法

上記の問題は、短い基本的な呼応表現の形式は、それを使用した表現に完全に含まれて出現することが多いため起こる現象である。長単位の表現に含まれる基本的表現を抽出するため、表現が完全な部分文字列であっても、他の部分にも出現する表現であれば収集する**弱抑制型連鎖共起表現抽出法**を提案する。

本方式は、従来の強抑制型連鎖共起表現抽出方式において、完全な部分文字列であるため収集対象外となった表現を、同じ表現で外部に頻度２以上で出現する文字列を収集する際に復活させ頻度の集計に加えることによって実現することが出来る。

本方式を日経産業新聞３カ月文（９７２万字）に対して適用し、連鎖共起表現を求めた結果を表１、表２に示す。基本的な表現の収集を目的とするため、表１、表２とも収集対象としたのは、漢字（漢数字を除く）、平仮名、片仮名に限り、また平仮名を１文字以上含んでいる２文字以上の表現である。

表１は、抽出対象表現の文字長による影響を示したものである。出現頻度が２以上の表現について示してあるので強抑制型と弱抑制型の抽出方式では文字列の種類数は同一なっている。延べ出現回数を比較すると長い文字列に対しては大きな差はないが、２文字以上、５文字以上の部分で、それぞれ約５倍、２倍となっており、弱抑制型ではより基本的な表現の抽出回数が大幅に増えていることが分かる。

表１　文字列長と出現頻度の関係

抽出対象表現の文字長	強抑制型		弱抑制型		無抑制型（長尾・森の方法）	
	文字列の種類数	延べ出現回数	文字列の種類数	延べ出現回数	文字列の種類数	延べ出現回数
２文字以上	825,262	2,390,016	825,262	11,345,019	3,540,200	27,298,173
５文字以上	486,710	1,263,629	486,710	2,194,356	2,273,629	8,721,156
１０文字以上	42,253	94,295	42,253	108,928	386,498	929,306
２０文字以上	1,204	2,487	1,204	2,562	42,475	87,751

表２は出現頻度の大きな表現をどれだけ収集できるかを示すデータである。弱抑制型の抽出方式では頻度２００以上の表現でも種類にして７０００、総数５７万という表現が抽出されている。また、抽出された表現の中には１万以上の出現頻度をもつものも多数出現している。しかしながら、頻度の大きな部分においてはその数が無抑制型の抽出方式に近くなってきており、離散共起表現への入力にあまり大きな数を足切りラインとすると、断片的な文字列が占める割合も高くなってしまい、精度が悪くなることが予測される。

表２　出現頻度

抽出対象表現の出現回数	強抑制型		弱抑制型		無抑制型（長尾・森の方法）	
	文字列の種類数	延べ出現回数	文字列の種類数	延べ出現回数	文字列の種類数	延べ出現回数
２回以上	825,262	2,390,016	825,262	11,345,019	3,540,200	27,298,173
５回以上	69,905	652,632	213,921	9,925,269	772,063	20,646,337
１０回以上	15,790	320,964	111,292	9,263,070	329,822	17,834,123
２０回以上	3,662	167,885	59,211	8,567,684	150,658	15,468,604
５０回以上	621	82,189	26,203	7,566,588	55,540	12,638,934
１００回以上	175	52,329	14,041	6,726,341	26,220	10,619,451
２００回以上	53	35,915	7,169	5,769,377	11,858	8,635,652

３．離散共起表現の抽出

前章で提案した手法によって得られた連鎖共起表現を要素として、離散共起表現を求めた結果を表３に示す。入力要素となる連鎖共起表現の出現頻度が大きく、対象データすべてを同時扱うことが困難であったため、全体を６分割し、各々に対して離散共起表現を抽出し、結果を集計した。連鎖共起表現の収集においては全体を１つとして扱っているため、収集対象となる表現は各分割データとも共通となっており、また、今

回は頻度の大きな表現のみを対象とするため、分割による大きな影響はないと考えられる。表中の最大ファイル量は6分割したデータ中、最大となったデータの値を示している。

離散共起表現の抽出方式としては、基本的な無抑制型の抽出方式に加え、同一文中に同一の表現が複数回出現する場合に、表現の不要な組を絞り込む弱抑制型抽出方式と強抑制型抽出方式が提案されている[3]。

連鎖共起表現の抽出に弱抑制型を使用した場合、短い単位で抽出される表現の頻度が高いため、同一文中に同じ表現が繰り返し現れることも多いと考えられるため、ここでは強抑制型の離散共起表現抽出方式を採用した。

離散共起表現抽出への入力データとしては、出現頻度８０以上の連鎖共起表現を対象とし、出力は頻度２０以上の離散共起表現とした。

表3に示すとおり、3要素以上の離散共起表現で、高い頻度で出現するものが多数抽出されており、より基本的で、高い頻度を持つ表現を抽出するという目的通りの結果を得ることが出来た。

表３　離散型共起表現収集結果（頻度２０以上）

	2要素の組	3要素の組	4要素の組	5要素の組	6要素の組
延べ組数	12,828,209	4,337,884	40,134	1,250	24
異なり組数	72,816	43,250	941	42	1
最大ファイル量	718.4MB	1658.4MB	496.6MB	7.4MB	0.6MB

4．考察

実験結果に基づき、本方式における抽出表現の特徴について説明する。本方式は、より基本的な単位での表現の抽出を目的としたもので、高い頻度で出現する離散共起表現を抽出することが可能となる。本手法により抽出された離散共起表現の例を表4、表5に示す。

表４　抽出例（文字列長順）

筋が～日明らかにしたところによると(45)
日明らかにしたところによると～する(70)
筋が～明らかにしたところによると～する(23)
の衆院予算委員会で～について～した(24)
相は～の参院～委員会で～について～した(20)

表５　抽出例（頻度順）

する～てい(9179), から～てい(8136), など～てい(8042), する～いる(7806), して～する(7667)
から～する～てい(1128), して～する～した(1109),して～こと～した(1089)

文字列長順にソートした結果では、文型がほぼ正しくとられていることが分かる。しかしながら、抽出されたデータ全体では長単位でとられている語が少なくなってしまっている。今回の実験においては、計算量を抑えるために連鎖型共起表現の入力足切りを出現回数８０と高めに設定したため、元々出現頻度のそれほど大きくない長単位の表現が落とされてしまったためである。

表6に入力足切りとなってしまった長単位の表現例を示す。頻度２０から７９で足切りとなってしまった長単位の表現は約８７００種類にのぼり、特に文末表現や、固有名詞を含むものなど有効な表現が多いため、これらが切り捨てられてしまう状況には問題がある。この問題を回避するためには、入力足切りを単純に頻度だけで行うのではなく、出現頻度と文字長などを組み合わせた制限方法をとることが必要になると思われる。

機械翻訳などを行うために用例を収集し、パターン化していく場合は、表現をなるべく長単位でとらえた方がパターン化が行いやすいと考えられるため、入力足切りの方法に関してはさらに検討していく必要がある。

表6　入力足切りとなった長単位の表現例

あることを明らかにした(52), あるからだ(65), あがっている(53), あげられる(57), あまりにも(79), あらかじめ(78), ありそうだ(68), ある程度の(73), いずれにしても(51), との見方が強い(64), は考えられない(62), インドネシアの(55), トヨタ自動車工業と(34)

また頻度順にソートした結果においては、完全には基本的文型とはいえないものが混ざってしまっているが、出現頻度は数千を越える表現が数多く抽出されている。しかしながら、２要素、３要素のレベルでは、全体的に抽出される異なり組数が多いため、まだ人手で文型の収集を行うには困難な状況である。こちらに関しては、連鎖共起表現の抽出の精度を向上させることが必要となる。部分文字列の収集において、現在は部分文字列でない表現が外部に２以上あれば収集対象としているが、外部に現れる回数と内部に現れる回数の比率が一定の値を超えた場合にのみ、収集対象とするなどの方法で精度の向上が可能だと考えられる。

また、連鎖共起表現の抽出精度向上の手法としては、抽出文字列のエントロピーを利用する方法[4][5]などが提案されているが、エントロピーを求める計算量が大きいなどの課題がある。基本的な文型を抽出していく方法についても、今後検討を進めていきたい。

５．まとめ

本論文では、連鎖共起表現の抽出に弱い抑制を掛けることにより、より基本的で、高い出現頻度を持つ表現を抽出する手法を提案した。また離散共起表現抽出を、本手法により抽出された連鎖共起表現を基に行うことで、出現頻度の高い表現を抽出することが出来ることを示した。この手法を９７２万字の新聞記事データに適用した実験においては、出現頻度が１０００を越えるような離散共起表現を数多く抽出することが出来、本手法が有効であることを確認した。

今後は、共起表現抽出の際の足切り条件の見直しや、部分文字列の収集方法の変更などを行いデータ抽出の精度の向上をはかるとともに、抽出したデータの応用方法など、より具体的な利用に際しての検討を進めていく。

参考文献

[1] 長尾，森：New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese, COLING'94, pp.611-615

[2] 池原，白井，河岡：「大規模日本語コーパスからの連鎖型および離散型の共起表現の自動抽出法」，情報処理学会論文誌, Vol.36, No.11,pp.2584-2596(1995)

[3] 池原，白井：「大規模日本語データからの離散型共起表現の自動抽出」，平成７年電気関係学会関西支部連合大会, G14-2

[4] 浦谷：「ニュース原稿データベースからの表現パターンの抽出」，情報処理学会第５０回全国大会, 1R-8(1995)

[5] 下畑，杉尾，永田：「隣接文字の分散値を用いた定型表現の自動抽出」，情報処理学会自然言語処理研究会報告110-11,pp71-77(1995)