様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　５月３１日現在

機関番号：13102
研究種目：基盤研究(C)
研究期間：2009～2011
課題番号：21500133
研究課題名（和文）　構文片言語単位の提案と統計的主観表現処理における有効性検証

研究課題名（英文）　Proposal of syntactic piece: its idea and application to statistical natural language processing tasks

研究代表者
山本 和英（YAMAMOTO KAZUHIDE）
長岡技術科学大学・工学部・准教授
研究者番号：40359708

研究成果の概要（和文）：
自然言語処理の研究における新しい処理単位として構文片という概念を提案し、その有効性を検証した。構文片は構文情報を持った最小の言語単位として考案され、単語集合やn-gram にはない様々な利点を持つ。また複雑な処理を要しないことから様々な統計処理との親和性が高い。本課題においてはいくつかの問題点を解決した上で構文片の定義を確定させ、さらに評判分析・自動要約という２タスクにおいて従来の処理単位よりも有効であることを確認した。

研究成果の概要（英文）：
I have proposed "syntactic piece" as a language processing unit of Japanese. Syntactic piece is the smallest unit that keeps syntactic relation. I have shown in this grant that the syntactic piece outperforms other units such as n-gram and bag-of-words, in the task of sentiment analysis and automatic summarization.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,400,000	420,000	1,820,000
2010 年度	900,000	270,000	1,170,000
2011 年度	1,100,000	330,000	1,430,000
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

研究分野：自然言語処理
科研費の分科・細目：情報学・知識情報学
キーワード：構文片、形態素、言語処理単位、n-gram、統計的言語処理、自動要約、評判分析

１．研究開始当初の背景

自然言語処理の研究をする際、その目的に適した処理単位を用いることは当然である。現在、使用される処理単位は対象にする問題によって違うが、主な処理単位として用いられるものはほとんどの場合単語集合、もしくは単語や文字の n-gram である。しかし、これらの処理単位はそれぞれ問題を抱えている。例として、言語表現に positive・negative の極性を与える問題を考える。単語集合を用いた場合、語義曖昧性をもつ単語はその曖昧性にまったく対応できず、すべて同じ極性の表現として扱うこととなる。たとえば、「与える」という単語は、「プレゼントを与える」という文脈では positive な表現だと言えるが、「誤解を与える」という文脈では negative な表現である。しかし、単語集合ではこれら

の区別をつけることはできない。一方、n-gram、上記の例でいえば単語 3-gram などを用いれば「プレゼント, を, 与える」という表現に対して極性を付与でき、この語義曖昧性の解消を期待できる。しかし、「を, 与える, と」「よう, だ, ね」のように意味をもつとは言えない表現が生成されてしまうことがあり、無駄が多くなる。

このような問題を解決するための処理単位として提案されたのが構文片である。構文片は、意味のある表現を処理単位とすることを目的としており、現状は構文解析の結果から係り受け情報を取得し、その結果から抽出された修飾節と被修飾節の対を構文片として使用している。構文片と単語集合と比較すると、構文片は修飾節を含むため、語義曖昧性による問題の多くは解消される。また n-gram と比較すると、構文片は構文解析を行った結果から生成されるため、少なくとも文法的な意味を抽出された表現に持たせることができ、無駄な表現は少なくなる。

しかし、構文片にはいくつかの問題がある。まず、構文片は文節の対で形成されるため、文字列も長くなりがちである。そのため、たとえば辞書として構文片を扱う場合に照応がとりにくくなる（過疎性問題）。また現在の構文解析ツールを用いると、品詞の分類の影響から「こと-が⇒ある」のように意味を持たない文節対も１つの表現として抽出されてしまう。これらの問題の影響は大きく、構文片はまだ実用的に扱える処理単位とは言い難い。

青木らは構文片の提案を行い、それを評判分析手法へ適用させて実験を行った。この研究では構文片の構文構造を保持した特徴に着目し、構文片を素性とした評価表現辞書を作成・拡張した。そしてその辞書を用いることで文単位の評判分析を行った。その結果、単語 2-gram や単語 3-gram では抽出できない評価表現を抽出することができ、評判分析における構文片の有効性を示している。一方で他の処理単位に比べ再現率が下回る結果となり、これを問題点として挙げている。これは、構文片という文字列の長い要素を処理単位とすることで対応できる評価表現が減少し、分類できなかった文が多く存在することが主な原因である。本稿では、構文片の汎化手法を複数提案することで、このときの構文片よりも再現率を向上させること期待する。

藤村らは評判分析に使用するときの評価表現の問題として、「大きい」といった文脈によって肯定・否定の意味が変わるような単語の指摘をしている。そしてその問題を改善するため、文を構成する主要な単語のみを用いた文節の n-gram を素性として提案した。文節 n-gram を用いることで、共起性の高い連続した句を素性として採用できるとしている。この動機付けは、「単語単位では語義曖昧性に対応できない」という、構文片を使用する理由の１つと同様である。そしてこれを素性とした評判分析を行い、その結果単語レベルの素性よりも高い効果を示している。しかし文節 n-gram では、たとえば「画面が /とても/ 大きい」のような表現があるとき、評判表現としては「画面が/ 大きい」のみで十分だが、「とても」を含んだ文節 3-gram を取得しなければ評判表現として扱うことができない。つまり取得した素性に冗長な表現が混じるおそれがある。一方構文片では、この例においても「画面が⇒大きい」「とても⇒大きい」の２つを素性として取得することができる。また文節も 2-gram までしか取得しないため、文節 n-gram よりも冗長性の少ない素性として扱うことができる。

２．研究の目的

構文片は以下の特徴を持つ。

● 単純な係り受け関係のみを用いているため、n-gram のように誰でも容易に抽出が可能である。統計情報も取りやすく扱いやすい単位である。

● 部分的な構文構造を保持している。そのため構文片の一片一片が意味を持った単位であり、単語連接では抽出が困難な表現も抽出することが可能である。

● 語順が自由な言語にも対応できる。日本語や韓国語のような比較的語順が自由な言語では単語連接の組合せは膨大となる。そのような言語では n-gram モデルはうまく機能しないと考えられる。構文構造を考慮している構文片は語順が自由な言語にも対応できる。

● 意味のまとまりとして取り扱うことができる。例えば"肌-が⇒合う"のような慣用句でもひとつの単位として扱うことができる。他の処理単位ではそれらの問題を無視するか、外部から慣用句の辞書を作ることで回避していた。構文片では慣用句と同様に扱うことができるため、あらかじめ辞書を用意したり、慣用句の同定を行ったりをする必要がない。

● 語義曖昧性に容易に対応でき、その誤りも少ない。例えば、「引く」という単語は「人目-を⇒引く」「ドア-を⇒引く」のように、共起する単語によって意味が大きく変化する。単語集合ではこの用途の違いを判別することはできない。単語 n-gram を用いれば抽出することができるが、その代わり「を, 引く, の」のような意味をまったくもたない無駄な要素も同時に抽出してしまう。しかし、構文片なら係り受け情報により無駄な要素を取得することなく曖昧性の問題を解決できる。

● 多くの場合文の復元が可能である。構文片はその名の通り構文構造の一片であり、構文片を組み合わせることで文を復元できる。

その一方で、現状の構文片には大まかに分類して以下の２点の問題がある。

（1）抽出される要素が他の処理単位に比べ冗長になるため、スパースネスの問題が発生する
（2）構文解析の結果をそのまま使用すると、意味を持たない文節対も抽出されてしまう

(1)の問題は、構文片が文節の対から生成されることが原因である。このような生成を行うと、当然得られる表現は冗長となる。冗長のまま構文片を扱うと、ほとんど同じ意味をもつ表現にも関わらず別々のものとして扱われ、結果的にスパースネスになってしまう。また(2)の問題の意味を持たない文節対とは、「こと-が⇒ある」や「なる-と⇒思う」といったものを指す。現状の構文片ではこのような文節対が構文片として扱われているが、意味が通じないことが見てとれる。これは、「意味のある要素」を処理単位として扱うことを目的としている構文片として適切でない。本論文では、これらの問題を解決するための手法を提案する。
　まず(1) の問題点の解決法として、以下の３つの汎化手法を提案する。

- 同類表現の統一
- 上位語への換言
- 機能動詞のラベル付与

　また、(2) で挙げた意味を持たない文節対が生成される問題については、「こと」や「ある」などの内容語に分類されている単語が、実質的に機能的な役割を果たしていることが原因だと考えた。そのため、これらの単語を機能語として扱う規則を用意することでこの問題の解決を試みる。

３．研究の方法

（1）同類表現の統一
　構文片には、その構成が原因となり、ほとんど同じ意味を指すにも関わらず、表現としては別のものとして扱われてしまうものがある。具体例を例１ に示す。

例 1
ケーキ-が⇒おいしい/ おいしい⇒ケーキ

これら２つの表現は、構造は違っていてもそれぞれが示す事象の意味は類似していると言える。このような特徴を持つ表現のことを、本論文では「同類表現」と呼ぶ。この表現を現状のまま扱うと、たとえば構文片で統計を得る場合に「ケーキのおいしさ」を表す表現を厳密に計測することができなくなる。そこで、同類表現をすべて同じ意味を持つ要素として扱えるように統一する。
　統一のために、同類表現を以下のように定義する。

1. 構文片が含む内容語はすべて同じである。
2. 構文片の分類が格フレーム、形容詞修飾、および動詞修飾である。

　この手法には構文片を汎化できる以外にも利点がある。それは表現の意味がほとんど落ちないことである。一般的に汎化を行う場合には表現の一部または全てを換言することで似た表現をまとめることが多い。そのため、汎化の手法次第では本来の意味がほとんど失われてしまったり、換言自体が失敗して意味が違うものに変わってしまうおそれがある。
　一方この手法では単語などを言い換える等のことはしておらず、本当に近い意味の表現のみを対象にすることが可能となっている。そのため汎化手法における懸念要素のひとつである適合率の低下が他の手法よりも軽減された汎化を行うことが期待できる。

(2) 上位語への換言
　シソーラスの上位下位概念を用いて単語を上位語へ換言する。シソーラスとは、言葉を類義語、上位・下位概念などの観点において分類した辞書のことで、同義語や多義語を扱うとき、特に換言の分野などで多く利用されている。例えば、上田らは依存構造の部分木を用いた複数文書要約を生成する問題において、同義な部分木を共通の部分木として扱うためにシソーラスを用いて類義語，上位語を換言する手法を用いている。たとえば、「チワワ」という単語を汎化するとする。この場合、

チワワ→犬→哺乳類→動物→…

のように上の階層にあたる単語に換言していく。
　前述したように、シソーラスを用いた同義語の汎化は、汎化を行う手法としては一般的なものである。しかし、換言していく階層の上限を誤ると、表現本来の意味が失われてしまうおそれがある。たとえば、「犬」「桜」という単語があり、これらを単語集合として扱うとする。この２つの単語を上位語に換言していった場合、「有機物」という単語以上の階層では同じ単語として扱われることにな

る。しかし、目的によっては両者を「動物」のカテゴリ、「植物」のカテゴリとしてそれぞれ区別させないと精度が落ちる場合も考えられる。そのため、目的やカテゴリによって汎化させる階層の上限を設定し、必要以上の汎化を防ぐようにしておく必要がある。一方、構文片は構文解析の結果から形成される単位であるため、文脈情報を保持している。そのため、構文片の中に含まれる単語を相当上の階層の単語に換言した場合でも、他の付随された文脈情報により単語の過剰な汎化を抑えることが可能となる。先ほどの例に当てはめると、

犬-が⇒吠える　→　[有機物]-が⇒吠える
桜-が⇒咲く　→　[有機物]-が⇒咲く

のように換言されることになる。「犬」「桜」の両方とも「有機物」という単語に換言されているが、それぞれに被修飾節を付与させることで両者を混同せずに扱えている。加えて、「犬」と「咲く」、「桜」と「吠える」が係り受け関係を持つことは考えにくい。ほとんどの場合「犬」と「桜」が同一の表現として扱わずに過剰な汎化を抑えることが期待できる。そのため、本手法ではカテゴリなどに応じて階層の上限を指定はせず、すべての換言を一律に扱うことができる。本研究では、再現率の向上を目的としているため、ルート(一番上の階層）から２番目まで汎化させることとした。

本手法では、シソーラスを EDR 電子化辞書内にある概念体系辞書を用いることとした。また汎化させる対象は構文片内における名詞・動詞を換言対象とした。

(3) 機能動詞のラベル付与

村上は、機能動詞を「実質的な意味を名詞にあずけて、みずからはもっぱら文法的な機能をはたす動詞」と定義している。たとえば「影響を受ける」という表現では、"受ける"という動詞本来の意味は薄れ、「影響される」と同じ意味を持つようになっている。この "与える"のような動詞が機能動詞に相当する。この特徴を利用して、機能動詞を含む構文片に対してラベルを付与する。そして同じラベルを持ち、かつその他の内容語が共通である構文片を同じ表現として汎化させる。

この手法はラベル付与による汎化だけでなく、構文片の異なり数を減少させることができる。例えば前述の例でも述べたように、「影響を受ける」と「影響させる」は同じ意味をもつ表現である。しかし構文片では「影響-を⇒受ける」で 1 つの要素として扱うことになる。このことから、１つの構文片を別の構文片の１文節に吸収させることが可能となり、異なり数の減少につながる。

機能動詞のラベル付与のため、村上の例を参考に人手で機能動詞を収集した。収集対象は基本的に動詞であるが、助詞との組み合わせによって機能動詞の働きをするものも存在するため、“助詞+動詞” の形で取得した機能動詞もいくつか存在する。また集まった機能動詞は態表現・相表現の分類に基づいてラベルを付与した。複数の分類を跨ぐことのある機能動詞も存在したが、これは今回収集対象外とした。

以下にそのリストを示す。

● 態表現
能動態(影響する)
受動態(影響される)
使役態(影響させる)
使役受動態(影響させられる)
交互態(影響しあう)
● 相表現
起動相(影響しだす)
終結相(影響した)
実現相(影響させる)
継続相(影響し続ける)
反復相(影響を重ねる)
強意相(影響を強める)
緩和相(影響を洩らす)

機能動詞を対象とした研究として、藤田らの研究がある。藤田らは、語彙概念構造に基づいた言い換え生成モデルを機能動詞構文の言い換えという問題に適用している。機能動詞の中でも態表現（「する」,「される」,「させる」）に対応するに対象を絞り言い換えを行った結果、言語モデルを用いたベースラインよりも高い性能を示すことができている。一方で、「時間"に"制限がある」という文を「時間"に"制限する」と誤って言い換えを行ったりと（「に」が「を」に言い換えられるべき）、いくつかの問題点を提示している。

本手法では、機能動詞を換言するわけではなく、機能動詞にあたる箇所にラベルを付与し、そのラベルを元に同じ表現を同一に扱う。そのため藤田らが挙げた問題点を考慮する必要なく機能動詞を扱うことができる。

(4)形式的内容語の結合

従来の構文片は修飾節と被修飾節の対であり、構文解析結果から係り受け関係を持つ文節の対を取得することで抽出している。しかしこの方法で抽出すると、意味を持たない文節対を取得することがある。例えば、「とても満足することができる」という文を構文解析すると
[1] とても⇒満足する
[2] 満足する⇒こと
[3] ことが⇒できる
の 3 つの文節対が抽出される。しかし[2]の

文節対は単体で１つの文節である。また、[3]の文節対は意味がまったく通じない表現となっている。これは「こと」という単語が形態素解析の分類の上では内容語に分類されているために発生する。ここでの問題点は、「こと」という単語はそれ単体では意味を持たず、実質的には機能的表現に分類されるべき単語ということである。本研究ではこのような単語を人手で収集し、「形式的内容語」と定義した。

形式的内容語を含んだ文節対を構文片とすることは、「意味のある要素を処理単位とする」という本来の目的とは外れてしまう。そこで形式的内容語を含む文節は、直前の内容語に対する機能表現として扱う。

上記の例にあてはめると、「満足すること」をひとつの名詞句として扱い、その修飾先を「できる」にする。つまり最終的には

[1] とても⇒満足する
[2] 満足すること-が⇒できる

という文節対に整形することで意味のない文節対を生成させることなく、本来の目的に沿った構文片を抽出することができる。本研究では以下の単語を人手で収集し、形式的内容語として扱うことにした。

４．研究成果

(1)評判分析への適用結果

問題点を改良した構文片の有効性を調査するため、評判分析に適用させる。評判分析とは、Web 掲示板や Weblog などのインターネット上に多く存在するある対象への評価や意見を自動的に得るための技術である。この技術により、対象となっている商品などの提供元などは効率的に評価情報を得ることができる。

本研究では評判分析の対象は文とし、１文を肯定・否定・その他に分類することとした。分類手法は青木らの手法を基に実装する。まずは、構文片自体を評価表現として取り扱うために構文片に極性を付与する。極性とは、表現が肯定・否定のどちらに属するかを示すものである。人手で用意した肯定・否定に分類した教師データを用いて、構文片の出現比率を計算する。肯定文に多く出現する構文片は肯定表現、否定文に多く出現する構文片は否定表現として抽出する。次に評価表現の種辞書を作成する。極性の付与された構文片を収集し種辞書とする。しかし種辞書だけでは教師データ中の評価表現しか扱うことができず、その教師データは人手で収集するためその規模に不安が残る。そこで大規模コーパスを利用し、種辞書を拡張することで辞書の規模を大きくする。最後に、用意された辞書を用いて評判文を肯定・否定・その他に分類する。ここでの「その他」は、評判を示さない文のことを指す。

教師データとして、人手により分類した肯定 1,966 文・否定 1,019 文の計 2,985 文を用意した。辞書拡張用の大規模コーパスには約 31.5 万文を用意した。教師データ・大規模コーパスともに、Yahoo!API を利用して取得した Yahoo!ショッピングレビューから作成した。

そして教師データを 5 分割し、１つをテストデータ、残りを学習データとして評価を行った。構文解析には構文解析器 Cabocha を用いた。実験手法には、各種提案手法を独立して行った場合と、ベースラインとして提案手法を使用しない従来の構文片を用いた。

文分類の結果を下記に示す。この結果から、機能動詞のラベル付与以外の手法では再現率・適合率どちらかは上回る結果を得ることができた。特に同類表現の統一、および動詞の上位語への換言ではどちらの精度も向上することを確認できた。

処理単位	再現率(%)	適合率(%)
同類表現の統一	49.8	77.1
名詞の上位語への換言	54.4	72.6
動詞の上位語への換言	51.5	76.2
名詞・動詞の上位語への換言	59.4	73.6
機能動詞のラベル付与	48.2	75.5
形式的内容語の結合	44.6	77.3
従来の構文片(ベースライン)	48.2	75.5

(2) 自動要約への適用結果

評判分析の実験により、提案した汎化手法を使用しても人手を介する手法では他の処理単位を上回ることが難しいということがわかった。つまり人手で学習データを用意することが必要ないものを対象とすることで、この統計上の不利を軽減できるのではと予想する。

そこで、統計的手法を用いた自動要約に本手法を適用させる。自動要約の手法は、文中の各要素に重み付けをし、その総和を重要度として扱う手法を用いることにした。具体的な手法を以下に述べる。

要約の対象となるテキストとして新聞記事を、重み付けとして tf*idf を用いる。tf は文書 d における索引語 t の頻度のことを示す。idf は、df を「文書数 N 中に t が一回以上出現する文書の数」としたとき、以下の式で求められる。

$$idf(t) = log10\ (N / df(t))$$

tf とは、出現頻度(term frequency) のことを指し、そのテキスト内でより多く出現する表現（ここでは構文片）は重要であると仮定した重みである。また、idf とは逆文書出現頻度(inverse document frequency) のこと

であり、より多くの文書で出現する単語はそのテキスト内では特別な単語ではないと仮定した重みである。これらをかけあわせたものが tf*idf である。つまり 1 つの文書中に頻度の高い索引語が多ければ tf*idf の値は大きくなり、多くの文書に出現する索引語を含んでいれば小さくなる。

１記事内に存在する各文を tf*idf で重み付けし、文中に存在する構文片の重みの総和を文の重要度とする。そして要約率を満たすまで、重要度の高いものから順に文を抽出する。最後に、人間が用意した要約を正しい要約と仮定し、出力されたシステムの要約と比較することで精度を計算する。

df を計算するためのデータとして、日本経済新聞 2000 年の記事 1 年分を使用した。総記事数は 201,829 記事であった。実験データとして 100 記事の正解データを人手で作成した。自動要約の結果を下記に示す。結果から、従来の構文片よりもいくつかの提案手法で従来の構文片の精度を上回る結果が得られた。

処理単位精度	(%)
同類表現の統一	34.0
名詞の上位語換言	31.5
動詞の上位語換言	33.0
名詞・動詞の上位語換言	31.5
機能動詞のラベル付与	31.8
形式的内容語の結合	32.9
従来の構文片(ベースライン)	32.6

(3) 結論

本研究では単語集合や n-gram に代わる処理単位として提案されている構文片の問題点の改善のため、3 つの汎化手法と、今まで抽出されていた意味を持たない要素を適切な形に整形する手法を提案した。そして改良を行った構文片の有効性を検証するため、評判分析および自動要約に適用させた。

まず評判分析では、人手で用意した教師データから得た統計情報を用いて作成した辞書を元に、入力文を肯定・否定に分類する実験を行った。その結果従来の構文片よりも全体的に精度がに向上し、本手法の有効性を検証することができた。しかし一方で単語 2-gram や 3-gram よりも大きく再現率が劣る結果となった。また適合率も若干ながら劣っていた。つまり、人手で用意した教師データを用いる小規模なシステムにおいて、本手法は n-gram より良い結果を得られないということを確認した。

次に、自動要約では新聞記事１年分を対象に、統計的手法である tf*idf を利用した重要文抽出の実験を行った。その結果、従来の構文片よりも高い精度で重要文抽出を行うことができた。また、単語集合や単語 n-gram と比較しても、これらより高い結果を得ることができた。これにより、少なくとも新聞記事１年分を対象にすれば、構文片は他の処理単位よりも優れた結果を得られることを証明できた。

以上より、本研究では評判分析と自動要約、２つの自然言語処理技術において従来の構文片よりも精度の高い構文片の実装と、大規模データ上における構文片の有効性を証明した。

今後の課題として、さらに再現率を向上できるような手法を考案し、小規模な学習データでも高い精度を得られるようにする必要性がある。そのための方法として、今回提案した汎化手法を組み合わせることが考えられる。また、汎化手法の１つとして提案した機能動詞のラベル付与においては、評判分析・自動要約どちらの実験でも効果が得られなかったため、適切な機能動詞の選別を行うことで良い精度を得られるようにする必要性がある。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計２件）

瀧川和樹, 山本和英.
構文片の改良と評判分析への適用.
言語処理学会第 17 回年次大会,
A2-5, pp.111-114, 2011.

Kazuki Takigawa and Kazuhide Yamamoto.
Syntactic Piece : Idea, Purpose and Application to Sentiment Analysis.
Proceedings of 7th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE 2011), pp.401-404, 2011.

〔その他〕

ホームページ等

http://www.jnlp.org/

６．研究組織

(1)研究代表者

山本 和英（YAMAMOTO KAZUHIDE）
長岡技術科学大学・工学部・准教授
研究者番号：40359708

(2)研究分担者

なし

(3)連携研究者

なし