文章一括処理による係り受け関係の解析

佐々木美樹　　坂本仁
沖電気工業（株）関西総合研究所ＡＩプロジェクト

1　はじめに

自然言語を機械的に処理する場合、文脈を考慮しなければ正しい係り先の決定が行なえないような係り受け関係が少なくない。しかし、従来は、係り受け関係を一文毎に個別に解析するだけであり、解析した結果を他の係り受け関係の処理に利用していない。このため、人間が読めば文章内に明らかな根拠があるのに誤って解析したり、同様の係り受け関係に対して正しく解析する場合と誤って解析する場合が混在したりするなどの問題がある。本発表では、文章全体から抽出した係り受け情報を利用した係り受け関係の解析を試みる。特に、再現性が高く係り受け情報が多く得られる語句単位の係り受け関係として、日本語の名詞句の係り受けと、英語のing-formの係り受けとを例として、文章を一括処理した係り受け関係の解析について述べる。

2　従来の手法の問題点

文章一括処理を行なわない場合は、係り受け関係を一文毎に個別に解析するだけであり、推定した結果を他の係り受け関係の推定に反映しない。例えば、日本語の名詞句では、「アメリカの日本への反発は．．．」という文においては、この文だけで「アメリカ」の係り先は「日本」ではなく「反発」であると意味から推定が可能であるが、「アメリカの業界への反発は．．．」という文においては、「アメリカ」の係り先が「反発」であると推定する手がかりはない。同じ文章内の「アメリカの反発は強く．．．」という文では「アメリカ」の係り先が「反発」であっても、それが手がかりになっていなかった。

また、英語では、辞書の情報により品詞を推定した後に文の解析を行なっているため、文章によって語の品詞や用法が変化することに対応できない。例えば、名詞句「radioactivity monitoring system」「technique underlying Ethernet」はどちらも「名詞＋ing-form＋名詞」の形をしているが、「radioactivity monitoring system」は「放射能監視システム」と訳すのが正しく「monitoring」は前置修飾であり、「technique underlying Ethernet」は「イーサネットの基礎となる技術」と訳すのが正しく「underlying」は後置修飾である。これらの修飾方向は意味から推定が可能であっても、この部分を構文解析によって正しく決定することは多品詞解消が正確に行なわれているという前提が必要であり困難であった。

文章から情報を抽出し曖昧性を解消する試みとして、構文解析の出力においてパックされている係り受け関係を展開して処理することによって前置詞句の曖昧性を解消する手法[1]や、入力表現との間のシソーラスに基づいた意味距離によって前置詞句の係り先のあいまい性を解消する手法[2]が提案されている。

本手法では、計算機で文章から抽出し処理できる表層的な情報だけで、係り受け解析の精度を向上させることを試みる。

3 解析方法

3.1 根拠となる関係の定義

3.1.1 日本語の名詞句

ある語の係り先は、何通りも考えられる。例えば、「アメリカの業界への反発」の場合、「アメリカ」の係り先は「業界」と「反発」が考えられる。しかし、「業界」は「反発」に係るほかない。この係り受けは語の意味を推測するまでもなく確実であると思われる。

よって、係り先の候補が1つしか存在しない、1番最後の係り受け関係を「確実な」係り受け関係とする。これを「根拠となる」関係とする。

更に、名詞の性質を定義する。根拠となる係り受け関係X→Yに対して、ある名詞が、Yである場合よりもXである場合が多いならば、Xを「前置性質」の名詞であるとする。逆に、Yである場合の方が多いならば、Yを「後置性質」の名詞であるとする。

3.1.2 英語のing-form

語が多品詞を持った状態で「根拠となる」関係の抽出を行なう。形態素解析結果が名詞・代名詞・数字・未知語のいずれかを持つ語を根拠抽出時の名詞とする。形態素解析結果が冠詞である語を根拠抽出時の冠詞とする。形態素解析結果が形容詞のみである語を根拠抽出時の形容詞とする。

「名詞＋ing-form＋名詞」の場合は、この部分だけでing-formの係り先を決定することが困難である。しかし、「形容詞＋ing-form＋名詞」の場合は、前の語が形容詞であるため、ing-formは後置修飾ではないと判断できる。「冠詞＋ing-form＋名詞」の場合でも同様である。そこで、「{冠詞｜形容詞}＋ing-form＋名詞」を「前置修飾の根拠となる」関係とする。同様に、「名詞＋ing-form＋形容詞」、「名詞＋ing-form＋冠詞」の場合は、この部分でのing-formは前置修飾ではないと判断できる。そこで、「名詞＋ing-form＋{冠詞｜形容詞}」を「後置修飾の根拠となる」関係とする。

更に、動詞の性質を定義する。あるing-formが、後置修飾の根拠となる関係である場合が多いならば、「後置修飾の性質」の動詞であるとする。前置修飾の根拠となる関係である場合が多いならば、「前置修飾の性質」の動詞であるとする。根拠となる関係により修飾方向が決定したing-formを「根拠となる動詞」とする。

3.2 根拠となる関係による係り受け関係の決定

3.2.1 日本語の名詞句

1番目の名詞をA、2番目の名詞をB、3番目の名詞をCとおいた「名詞＋名詞＋名詞」の3語組に対して、A→Bにのみ根拠となる係り受け関係がある場合はA→B、A→Cにのみ根拠となる係り受け関係がある場合はA→Cにする。A→BとA→Cの両方に根拠となる係り受け関係がある場合、または、A→BとA→Cの両方ともに根拠となる係り受け関係がない場合には、直後の語に係る係り受け関係を優先するように、A→Bにする。

名詞の性質を適用する場合は、Bが前置性質の名詞であるならば、A→BはBの性質上B→Cに比べ成立しにくい弱い関係であるといえるので、A→Cにする。Bが後置性質の名詞であるならば、A→BはBの性質上B→Cに比べ成立しやすい関係であるといえるので、A→Bにする。

3.2.2 英語の ing-form

根拠となる動詞の性質を適用する。ある ing-form が、後置修飾の性質の動詞であるならば、後置修飾にする。前置修飾の性質の動詞であるならば、前置修飾にする。根拠となる動詞がない場合には、前置修飾にする。

4 実験

4.1 日本語の名詞句

日本語では名詞句の係り受け関係として「名詞＋名詞＋名詞」の３語組を対象とした。１番目の名詞をＡ、２番目の名詞をＢ、３番目の名詞をＣとする。計算機関係の文章56000文から、曖昧さがある３語組約8000組を計算機処理により抽出し、４組に１組の割合で抽出した1980組のデータに対して、Ａ→Ｂの係り受けがあるかないか、Ａ→Ｃの係り受けがあるかないか、人手で正解を付与する作業を行なった。

適用した根拠は、以下の様である。

対象	根拠となる関係の種類	根拠となる関係を適用した数
Ａ→Ｂ	約 40000	509/1980
Ａ→Ｃ	約 40000	273/1980

係り先が一致した割合は、以下の様になった。

対象	係り受けがある割合	根拠となる関係を適用	根拠となる関係と名詞の性質を適用
Ａ→Ｂ	82 % (1624/1980)	95 % (483/509)	−
Ａ→Ｃ	33 % (648/1980)	51 % (140/273)	61 % (62/102)

4.2 英語の ing-form

英語では ing-form の係り受け関係として「名詞＋ing-form＋名詞」を対象とした。計算機関係のマニュアル、英字新聞、生産技術関係のレポートから、「名詞＋ing-form＋名詞」の係り受け関係を計算機処理により抽出し、ing-form が前置修飾か後置修飾である適当に抽出したデータに対して、人手で用法を付与する作業を行なった。

文章の分野	文章の語数	抽出した「名詞＋ing-form＋名詞」数(種類)	データ数
マニュアル	約 7688000	15992(3857)	277
新聞	約 1989000	4807(4335)	810
レポート	約 56000	255(211)	217

適用した根拠は以下の様である。

文章の分野	根拠となる関係の種類	根拠となる動詞の種類	動詞の性質を適用した数
マニュアル	3631	426	266/277
新聞	3911	889	727/810
レポート	134	63	143/217

修飾方向が一致した割合は、以下の様になった。

文章の分野	前置修飾である割合	動詞の性質を適用	動詞の性質を適用した正解率
マニュアル	78 % (216/277)	92 % (245/266)	91 % (253/277)
新聞	55 % (445/810)	84 % (611/727)	81 % (658/810)
レポート	83 % (180/217)	96 % (137/143)	87 % (189/217)

5 考察と課題

英語のing-formでは、どの分野の文章においても効果が確認された。動詞の性質を適用したデータが多く、動詞の性質の利用度は高い。根拠となる動詞が適用された割合も高く、効率がよい根拠といえる。文章毎のing-formの修飾方向の傾向に関係なく、修飾方向が一致した割合は向上し、文章間の差は減少しており、本手法は広く有効な手法であるといえる。特に、英語においては、語が多品詞を持った状態で係り受け関係を解析できるという点で、実用性が高い。今回は、「名詞＋ing-form＋名詞」において、ing-formが前置修飾でも後置修飾でもないデータは除いてあるが、今後は、根拠によってそれらを解析することができるかどうか試みていく。

日本語の名詞句では、根拠となる関係が適用された範囲で効果が確認された。実験で用いた3語組で、A→Cの係り受けがあるかないかを判定するという観点においては、A→Cの係り受けがある割合が低いため、すべてA→Cでないとする方がよい結果になるが、A→Cである係り受け関係を絞り込むという観点において、根拠となる関係と名詞の性質を導入した本手法は充分有効であるという見通しが得られた。今回は、根拠となる関係が2語対であるため、英語のing-formの動詞1語の場合に比べ、根拠となる関係を適用したデータが少なく全体の効果に寄与しないと思われるが、根拠となる関係の数を増やすことで適用する範囲を広げるようにして対処していきたい。A→BでありかつA→Cである場合が存在するため、修飾方向を解析する英語のing-formの場合に比べ、一方の解析結果をもう一方の解析に利用することができないという点で難しさがあった。今後は、A→BとA→Cを組み合わせた新たな根拠を探っていく。

6 おわりに

日本語の名詞句、英語のing-formを対象として、文章一括処理による係り受け関係の解析の手法を試みた。本手法は、意味情報や人手を必要とせず、表層的な情報によって文章全体で整合を取り、係り受け関係の正確さを向上させるなどの優れた点があるので、従来の係り受け関係解析を補強する手段として確立を目指したい。

今後は、語句レベルの解析から節や文レベルの解析に拡張していきたいと考えている。

参考文献

[1] 那須川：”文脈制約を利用した曖昧性解消”, 人工知能学会第7回全国大会. (1993).

[2] 隅田, 古瀬, 飯田：”英語前置詞句係り先の用例主導あいまい性解消”, 電子情報通信学会論文誌 Vol.J77-D-IINo.3 (1994).