並列構造の不正な統語解析結果を統計的に検出する

加藤鉱三（信州大），黒田航（NICT）

【1】問題

並列構造の解釈が言語処理の課題であることはよく知られている問題である。次の文を考えてみよう。日本語訳は，４つの市販翻訳ソフトのものである。

入力文(1)　The person responsible for your faults and the collapse of your relationship is you.

LV：　あなたの欠陥とあなたの関係の崩壊に関して責任がある人はあなたです。

The：　あなたの欠点のための担当者およびあなたの関係の崩壊は、あなたです。

Atlas：　あなたのせいに原因となる人とあなたの関係の崩壊はあなたです。

PC：　あなたの誤りとあなたの関係の崩壊に対して責任がある人は、あなたである。

○The person responsible for [your faults] and [the collapse of your relationship] is you.

×[The person responsible for your faults] and [the collapse of your relationship] is you.

(1)の出力で，並列構造については LV と PC は正しく分析しているが，The と Atlas は誤った統語解析結果を用いた訳文を出力している。一方，人間ならば，英語がある程度できる人なら，誤った解析をする可能性はない。それはなぜだろうか？

人間が誤った解析をしない理由は次のものであろう。

(2) a.　the person と the collapse が並列構造になっているとしたら，the collapse が述部 is you の主語であることになる

b.　しかし the collapse is you という主述構造は意味的にあり得ない

c.　よって(a)の解析はあり得ない

人間と同じ解析結果を機械に出させるには，次の２つの可能性があるだろう。

(3)　人間の行う処理をモデルとする

(4)　機械の苦手なことは避け，機械の得意なやり方をさせる

(3)は，少なくとも今問題にしている並列構造の解析に関しては，あまり現実的ではない。人間の処理の根幹部分は(2b)である。しかし(2b)の「意味的にあり得ない」という判断は正しいのだが，なぜあり得ないのかをはっきり説明しなくても人間には判断できる判断である。しかしこれを機械にやらせるためには，なぜあり得ないのかを明示的に示し，その判断の仕方を機械に教えなければならない。それをするためには，主語と述部の組み合わせに関する一般原則を明らかにする必要がある。しかし，この事例一つを取っても，the collapse is you がなぜ悪いのかを説明するのは実は難しい。更に，一般原則を明示化し得たとしても，その原則を適用するためには，少なくとも一つ一つの名詞や動詞が持つ情報を明示化しなければならない，という課題も次に控えている。並列構造の解析に関しては，機械が苦手とするやり方をしなければならない, という点において, (3)は(4)の対極にある。

(4)の方向では，いわゆる『統計翻訳』の手法が思い浮かぶかもしれない。しかし，少な

くとも並列構造の解析においては，それもあまり現実的なアプローチとは言えないだろう。それは次の理由による。

(5) [A and B] C または A [B and C] という並列構造では，ある部分がある要素によって共有されている，という点に特徴がある。共有されている，ということは，その部分は表面に現れていない，ということである。例えば she loves John and Bill では，she loves John and she loves Bill の下線部は表面に現れていない。

(6) 一方，統計翻訳は（前処理として構造解析をしないものである場合には），見えている部分同士の組み合わせと他言語での表現との対応を確率的に処理する，という性質のものであるため，表面に現れていない部分を読み取る，というようにはデザインされていない。

【2】提案

並列構造の特徴が(5)であるならば，その特徴を真正面から見据えた処理が当然有効であろう。並列構造の特徴が要素の共有にあり，かつ，表面に現れていない部分がある，ということであるならば，表面に現れていない部分を復元する，という方向性が考えられる。言語学的知見として，「省略されているものは必ず復元できるようになっている」ことが知られている。並列構造の場合には，表面に現れていない部分は共有部分であるため，それを(7)のように復元することは極めて容易である。(1)の場合は，(8)のように復元される。

(7) 共有部分の復元

[A and B] C という並列構造は，AC and BC と復元される

A [B and C] という並列構造は，AB and AC と復元される

(8) 入力文(1)での復元

responsible for your faults and the collapse of your relationship

⇒ [responsible for your faults]

and [responsible for the collapse of your relationship]

次に，このようにして復元された「見えない部分」を含む復元チャンクが適正なものであるかどうかを判定しなければならない。人間が判定するのであれば簡単であるのだが，(3)のアプローチが採れないことは上で確認したところである。機械に判定させるためには，『統計翻訳』以外の仕方で(4)の方向性を追求しなければならない。本発表ではそのやり方として(9)の方向性を提案したい。

(9) 復元チャンクの判定（英語の場合）

復元チャンクを web 検索し，その生起数で判定する

ただし検索フレーズは次の処理を施す（※網羅的なリストではない）

チェックする対象でない主語は外す

チェックする対象中の主語は，ヒトは you，モノは it に変える

主動詞以降は３語まで

共有部分は非共有部分に近い方から３語まで

(9)の処理には，統計的に有意となるような生起数を確保すること以外の理由はない。

このやり方では，(1)の復元チャンクを判定すべき検索フレーズは次のものとなり，それ

による Google 検索の結果は次のようになる。

(10)　正しい解析での検索フレーズ

[responsible for the collapse of your relationship]

→ “responsible for the collapse”

"responsible for the collapse" に一致する英語のページ 約 18,400,000 件

(1)で誤りとした統語解析についても同じことをしてみよう。

(11)　誤った解析での復元チャンクと検索フレーズ

×[The person responsible for your faults] and [the collapse of your relationship] is you.

共有部分

復元チャンク：　[the collapse of your relationship is you]

検索フレーズ：　[the collapse is your relationship is you]

→ “the collapse is you”

検索結果：　"the collapse is you" に一致する英語のページ 4 件

(11)の誤った解析結果から得られる検索フレーズは，復元チャンクの生起数が極端に低くなっている点に注意されたい。機械は意味を考えることはできないが，関連するものの生起数をカウントするのは得意である。このように，機械が得意なことを判定材料とする可能性を本発表では提案する。

【３】他の事例

入力文(12)　We thought that Cuisinart was a reputable brand and expected much better from this product.

LV：　我々は Cuisinart がこの製品から評判が良いブランドで、そしてずっと良く予想されたと思いました。

The：　私たちは、クイジナートは評判がよい商標でこの製品から多くをよりよく予期した、と思いました。

Atlas：　私たちは、Cuisinart が評判のよいブランドであり、予想された多くがこの製品から、より良いと思いました。

PC：　我々は、Cuisinart が評判の良いブランドで、非常によりよくこの製品に期待されたと思った。

○[We thought …] and [we expected …]

×We thought that [Cuisinart was a …]　and　[Cuisinart (was) expected …]

この事例では４つの翻訳ソフト全部が誤訳している。誤訳の原因は，and が並列させているものを誤って解析していることによるものである。この事例も先のものと同じように処理してみよう。

(13)　正しい解析

復元チャンク：　[we expected much better from this product]

検索フレーズ：　“we expected much better”

検索結果：　“ we expected much better ” に一致する英語のページ 約 825,000 件

(14)　誤った解析

復元チャンク： [Cuisinart expected much better from this product]
検索フレーズ： “it expected much better”
検索結果： “ it expected much better ” に一致する英語のページ 7 件

(14)は，日常的な言葉で言えば，「（ヒトは予測するが）モノは予測しない」ということである。それを機械に判定させるには，例えばこういうやり方がある，ということを提案したつもりである。

【４】今後の課題

次の文でも，４つの翻訳ソフトのうち the 以外が，mixed を，過去分詞ではなく誤って過去形と解析する誤訳を出力する。

入力文(15) Oil prices hovered below $78 a barrel Friday in Asia as investors eyed a volatile U.S. dollar and mixed economic data.

○[investors eyed a volatile U.S. dollar] and [investors eyed mixed economic data]

×[investors eyed a volatile U.S. dollar] and [investors mixed economic data]

この事例を上記のやり方で処理しようとすると，正しい解析から得られる検索フレーズは”you eyed mixed economic data”であるが，検索結果はゼロである。これは，eyed と mixed economic data の組み合わせが（正しいのだが）まれであることによる。この場合には，よりよい解決は，”the mixed economic data”を検索することにより mixed が過去形ではなく過去分詞であることを示すことによって得られる。このように，今後は事例のタイプ分けをし，タイプごとにどのような処理を機械にさせるかを考えていく必要がある。

また，これとは違う趣旨の課題もある。

入力文(16) We appreciate you and the job you are doing for us.

Atlas： 私たちはあなたが私たちのためにしているあなたと仕事に感謝します。

PC： 我々は、あなたが我々のためにしているあなたと仕事を評価する。

○We appreciate [you] and [the job you are doing for us].

×We appreciate [you] and [the job] you are doing for us.

この事例では Atlas と PC が誤った解析結果に基づく訳文を出している。これが誤っていることは，復元チャンク[we appreciate you you are doing for us]から得られる検索フレーズ”appreciate you you are”という不正な関係節の生起数が低ければ問題ないのであるが，実際には Google 検索で約 6,580,000 件という数字が出る。これらはほとんどが you と you の間に句読点を含むものであり，ここで検索すべき先行詞－関係節という連鎖ではない。フレーズ検索で句読点を認識していないために生じるこのような問題もある。

使用した翻訳ソフト

[LV]： 『LogoVista 2007 PRO』，LogoVista.

[The]： 『The 翻訳 2007 プレミアム』，東芝.

[Atlas]： 『ATLAS V14 翻訳スタンダード』，富士通.

[PC]： 『PC-Transer 翻訳 Studio 2008』，クロスランゲージ.