

文脈情報を利用した自然言語文における構造的曖昧性の解消

那 須 川 哲 哉[†]

自然言語処理で単語間の係り受け関係の曖昧性を解消する際に、文脈内の同じ語の振舞いを参照し実用的なレベルで文脈情報を利用することにより処理の精度を高める手法を提案する。さらに実験結果を通じてその有効性を示す。あらかじめシステムに辞書などの形で蓄えられた知識だけでなく、処理対象の文章から動的に抽出した情報を用いることで、知識の不足を補うことができる上に、文脈に適した解釈を優先させることができる。本手法は、文脈内で繰り返し出現する語句の係り受け情報を参照するという単純な処理で構成され、複雑な推論機構や文脈処理用の特別な知識に依存しないため実用性が高い。英文の計算機マニュアルを用いた実験では、前置詞句の係り受けの曖昧性解消において、距離的に最も近い候補に付ける場合の成功率69.7%に対して、本手法により82.6%の成功率が得られた。さらに、多品詞語や複合語、並列構造などの存在から複数の木構造が生成される場合に、正しい構造を選択する問題では、文脈を考慮しないヒューリスティックスでは74.5%であった成功率が、文脈の参照により89.1%へと向上した。

Structural Sentence Disambiguation Based on Discourse Information

TETSUYA NASUKAWA[†]

In natural language processing, domain knowledge plays important role in disambiguating input sentences; however, it is practically impossible to implement enough knowledge for covering various texts. On the other hand, interpretation of a sentence depends on its context, but the existing methods rely on hand-coded knowledge resources and complex inference mechanisms so that such methods are not yet adequate for practical systems. This paper describes a simple and robust method that improves accuracy of sentence disambiguation by using information extracted from other sentences in the same discourse through morphologically identical words. Since this method refers to basic information that consists of modifier-modifiee relationship derived from syntactic analysis, it is practical in terms of the amount of knowledge it presupposes and amount of computation it requires. In an experiment on computer manuals, accuracy of prepositional phrase attachment was improved to 82.6%, from 69.7% with default heuristics that attach the prepositional phrases to inner-most words, and accuracy of multiple parse selection was improved to 89.1% from 74.5%.

1. はじめに

自然言語処理では、入力文を解析する過程で、文中の各単語の意味（語義）や単語間の係り受け関係における曖昧性を解消する必要がある。そのために必要な知識や、その知識の構築法に関する様々な研究が存在する。ところが、多種多様な文章に対して十分な知識をあらかじめ用意しておくことは実質的に不可能であり、さらに解釈が文脈に依存する問題には、知識量に関わらず、単文単位の処理手法では対応できない。したがって文脈情報も重要な役割を果たすことが認識さ

れているが、従来の文脈処理手法は内容理解を意図した複雑な処理を前提とするものが多く、実用的なレベルには達していない。本論文では、実用的なレベルで文脈情報を利用する手法として、文脈内の同じ語^{*}の振舞いを参照することにより曖昧性解消の精度を高める手法を提案し、計算機マニュアルを中心とした技術文書における実験結果を通じて、その有効性を示す。

ある一定のテーマに関して記述された、一貫性のある文章においては、同じ語が繰り返し出現し、しかもその際には同じ語義で用いられる傾向がある。この性質を利用することで、文脈内で複数回出現している単

[†] 日本アイ・ビー・エム株式会社 東京基礎研究所
IBM Japan, Tokyo Research Laboratory

^{*} 辞書中で同じ見出しを持つ語。複数形や過去形など語形や語尾が変化していても同じ語とみなす。

語に関して、文脈内のどこか1カ所でも語義が決定できれば、語義が決定できない箇所でもその語義を選択することで、文脈全体における語義決定の精度を向上させることができる^{1),9)}。このような文脈内での一貫性を単語間の係り受け関係にも拡張して適用し、同じ文脈内で繰り返し出現する同じ語は同じような語と係り受けを結ぶように曖昧性を解消することで、文脈全体における解析精度を向上させようというのが本論文で提案する手法の基本的な考え方である。したがって、本手法は深い意味的な問題には取り組まず、完全な解析精度を追求するものではないが、文脈処理用の特別な知識や推論機構を用いずに自然言語処理の解析精度を向上させられる点で頑健性が高く、文脈依存の曖昧性の解消にも対応できるなど、実用性、発展性に富んでいる。

第2章で本論文が対象とする自然言語文の曖昧性の問題とそれに関する従来の研究の概要を示した後、第3章で文脈情報を用いた処理の概要を示し、第4章で具体的な処理の流れと実験結果を示す。

2. 自然言語文の曖昧性

自然言語文の解析においては、処理単位となる単語を切り出す形態素解析や多義語の語義決定など、様々な処理段階で曖昧性が発生するが、本論文では単語間の係り受け関係に関する構造的曖昧性に対象を限定する。また構造的曖昧性を、「係り受けの曖昧性」と「多品詞語や複合語、並列構造などにより複数の木構造を生成する構造的曖昧性」の2種類に分けて扱う。

2.1 係り受けの曖昧性

まず英文の具体例で係り受けの曖昧性の問題を示す。

(2.1) He saw a girl with a telescope.

という文の“with a telescope”は、“saw”に係る場合は「望遠鏡を使って（見る）」、“girl”に係る場合は「望遠鏡を持っている（少女）」というように、係り受けにより解釈が異なる。したがって、文の解釈を決定する過程で係り先を決める必要がある。(2.1)の例文は、文法的にも意味的にも複数の解釈が可能であり、曖昧性解消には文脈の参照が必要となるが、

(2.2) He saw a comet with a telescope.

の場合は、“a comet with a telescope”（望遠鏡を持つ彗星）という解釈が通常ありえないため、“with a telescope”は“comet”よりも“see”に係り易い^{*}という知識を利用して曖昧性を解消することができる。

このように、文法的な制約で曖昧性が解消できなくても、意味的制約から解消できる場合が多いため、このような意味的制約の知識を蓄えておき、その知識を用いて係り受けを決定する手法が、機械翻訳システムなどで一般的に用いられている。ところが、実際に汎用的な自然言語処理システムを構築し運用する上では、膨大な量の語句に関する係り受け関係の知識が必要であり、それを構築し整備するコストが大きな問題となる。したがって近年では、このコストを下げるため、既存の機械可読テキストデータから人手をかけずに自動的に知識を抽出し、それを用いて係り受けの曖昧性を解消する手法が盛んに研究されている。例えば、大量の文を処理した結果得られる依存構造を学習事例として半自動的に蓄積し、事例中に存在する係り受けパターンを優先する手法^{8),11),13)}や、大量の文における共起関係から特定の動詞や名詞に係り易い前置詞の知識を抽出し利用する手法²⁾、また、異なる言語間では係り受けの曖昧性の出現箇所が異なるという性質を利用して、二言語対訳コーパスから、曖昧性を排除した係り受けパターンの知識を抽出する手法¹⁵⁾をあげることができる。

このように、従来の曖昧性解消手法では、あらかじめ知識を構築しておくことが前提となっており、質の高い知識をいかに効率良く確保するかに研究の焦点が当てられてきた。しかし、完全な知識を構築するのは実質的に不可能であり、新しいテキストを処理する際には、必ず何らかの知識の追加が必要となるのが実情である。また、ある程度広い範囲の文章に対して十分高い解析精度を実現するためにどれだけの量の知識を集めればよいかの見通しも立っていない。さらに、(2.1)のように解釈が文脈に依存する場合は、知識がいくら存在しても上記の手法では解決できない。文脈に依存する問題を扱う研究としては、文脈を表現するためのスクリプト的な知識を利用し、対象世界に関する十分な知識を備えていることを前提とした手法³⁾や、すべての可能な世界モデルを構築し、新たな情報を読み込む過程で矛盾するモデルを棄却することにより意味を絞り込んでいく手法⁷⁾などが提案されているが、知識構築の困難性や処理の複雑性から、実験レベルにとどまっている。

2.2 多品詞語や複合語、並列構造などにより複数の木構造を生成する構造的曖昧性

構文解析の結果として得られる単語間の係り受け関係を木構造として表現すると、係り受けの曖昧性は、係り先の曖昧な句に対して係り先語句の候補を記述しておく形で、一つの木構造上で表現できる(図1(2))。

^{*} すなわち、“a comet with a telescope”よりも、“see ~ with a telescope”の解釈の方が優先される。

- (1) 「時は矢のように飛ぶ」
 flies <動詞>
 └ (SUBJ) — Time <名詞>
 (like) — arrow <名詞>
- (2) 「矢 [矢印] のような (ように) 蠅 [飛行] を計
 時して下さい」
 Time <動詞>
 └ (OBJ) — flies <名詞>
 ? (like) — arrow <名詞>
- (3) 「時蠅は矢 [矢印] を好む」
 like <動詞>
 └ (SUBJ) — (Time flies) <名詞>
 (OBJ) — arrow <名詞>
- (4) 「矢 [矢印] のような時蠅」
 (Time flies) <名詞>
 └ (like) — arrow <名詞>

図 1 複数解析構造の例

Fig. 1 Example of multiple parses.

ところが、多品詞語や複合語、並列構造などの存在により可能になる複数の解釈の曖昧性は、一つの木構造上では表現できず、係り受けの曖昧性解消とは異なる処理が必要となる場合がある。すなわち、係り受けの曖昧性の解消においては、決定した係り受けが他の語句の係り先候補を制限する場合がある以外には、同じ木構造中の他の部分の構造には影響しないが、ある多品詞語に対して異なる品詞を選択すると、他の語の品詞や係り受け関係も変化させなければならない場合がある。

例えば、

(2.3) Time flies like an arrow.

という文は、意味的な妥当性を考えなければ、文法的には主動詞を、“Time,” “flies,” “like” のいずれに解釈するか、あるいは全体を名詞句と解釈するかという曖昧性を含んでおり、解釈可能な依存構造をすべて表現するには、図 1 に示すような複数の木構造が必要となる[☆]。このように複数の解析構造が得られる場合、一般的には、各単語の品詞別出現頻度に関する知識^{☆☆}や、通常どのような文法構造がより多く出現しているかという知識に基づいて各解析構造の文法的妥当性を評価し、評価値の最も高いものを選択する手法が取られる。すなわち、文法的な構造の選好度に基づいて解析構造の各候補を評価し、最も妥当性の高いものを選択する。しかしながら、文の解釈は文脈に基づい

て決定されるべきであり、文脈を考慮しない手法では、出現頻度の低い文法表現は全く選択されなくなる危険性がある。

3. 文脈情報を利用した曖昧性解消

3.1 係り受けの曖昧性の解消

一つの文脈内では、同じ語が同じ意味で用いられる傾向にある^{1),9)}。☆☆ことから、係り受けに関しても、文脈内では同じ語が同じような語句と係り受けを結ぶと仮定すれば、係り先の曖昧な語に関して、文脈内の他の文で同じ語がどのような語と係り受け関係にあるかを調べ、それと同じような係り受けを優先することで、曖昧性が解消できる。

以下は 1 冊の計算機マニュアルから抽出した実際の文である。

(3.1) The system displays message EKC0246A on the MTO console.

(3.2) The messages are normally displayed on the MTO console.

(3.3) The messages are normally displayed by CICS on the MTO console.

(3.1) では“on the MTO console”の係り先が曖昧であり、“message EKC0246A”と“displays”の二つの候補が存在する。この曖昧性を解消する際に、(3.2)で“on the MTO console”が“displayed”に係っているという情報を参照し、文脈内の他の文と同じ係り受けを優先することで、係り先を“displays”に決定することができる。また、(3.1)と(3.3)は各々が係り受けの曖昧性を含んでいるが、両方の文で“on the MTO console”が同じ係り受け(動詞“display”との係り受け)を結ぶように、すなわち、文脈内での同じ語の係り受け関係に一貫性が生じるように処理することで、両方の曖昧性が解消される。

上記の例のように、計算機マニュアルのような技術文書では、同じような事柄の説明が何度も繰り返され、しかも、同じ事柄を再び記述する際には表現が多少変化する傾向にあるため、この変化による差分を利用することで曖昧性が解消できる。

3.2 多品詞語や複合語、並列構造などにより生成される複数の木構造の選択

この問題でも、係り受けの曖昧性解消と同様に、文脈内で同じ語が同じような語句と係り受けを結ぶことを前提とし、他の文と同じ係り受けを含む木構造を優

☆ その上、例えば、(2)の解釈において、“like an arrow”の係り先を“flies”にして「矢のような(蠅)」と解釈するか、“Time”にして「矢のように(計時する)」と解釈するかは「係り受けの曖昧性」も存在する。

☆☆ 例えば、“a”という語は、“once a day”における前置詞的な使われ方は稀で、大抵の場合、冠詞として使われる」という知識。

☆☆☆ 文脈内における語義の一貫性は、95%以上という非常に高い割合で成立することが報告されている。

先して選択することで曖昧性を解消する。

例えば (2.3),

Time flies like an arrow.

の複数の解釈から一つを選択する際に、もしも

The fly likes refrigerators.

のように, “fly” が名詞として用いられ, 動詞 “like” に主格として係っている文が同じ文脈内に存在すれば, それと同じ係り受け関係を含む解釈 (3) 「時虻は矢 [矢印] を好む」の構文木を優先する。

また, 連続する文は同じ文型を取ることが多い, すなわち, 操作法を説明する文章では命令文が連続し, 箇条書では名詞句が列挙されるというように, 連続する文では, 文法的構造に一貫性が成立し易いと仮定し, 前後の文と同じ文型の構文木を優先する。

4. システムの構築と実験

前章で示した曖昧性解消処理手法を英日機械翻訳システム Shalt2¹²⁾上で実現し, 計算機マニュアルの文章で実験を行った。本章では実際に用いたデータ構造および処理の流れを示した上で, 実験結果を示す。

4.1 文脈モデルのデータ構造

本手法では, 文脈の構成要素である各文に関して,

- 文の位置 (文脈内で何番目の文か)^{☆1}
- 文中の各語の形態情報 (品詞, 語の原形など)
- 文中の各語の依存関係 (どの語がどの語に, どのような関係^{☆2}を介して係っているか)

の情報を蓄積したものを文脈モデルとして利用する。以上の情報はすべて通常の構文解析結果から得られるため, 文脈処理用の特別な処理は全く行わずに, このモデルを構築することができる^{☆3}。基本的にはこれだけで十分であるが, 参照処理の効率を向上させるため, 同じ語の情報をまとめて保持する文脈フレームを構築する。すなわち, 入力文章を1文ずつ読み込み, 構文解析して文脈モデルを構築する過程で, 語単位で,

- どのような語に係っているか/係り得るか
 - どのような語に係られているか/係られ得るか
- という情報を蓄積していく。係り受けに関する文脈情報の具体例を表1に示す。表1には, 700文からなる計算機マニュアルの文章における動詞 “define” に関して, 係り受けを結ぶ語の品詞と接続関係ごとに, 係り先の語と (修飾要素として) 係る側の語, そこに与えられる選好度^{☆4}が記述されている。例えば, この文脈では代名詞 “you” が主格を介して “define” に係るパタンの選好度が30である。

4.2 曖昧性解消処理の流れ

曖昧性解消は, 以下の流れに従って行われる。

(1) 文脈情報の読み込み (文脈モデルの構築)

まず入力文章中のすべての文を1文ずつ構文解析し, 係り受けの曖昧性は解消せずに残したまま, その結果を文脈情報として蓄積していく。ただしその際, 1文につき一つの木構造の情報を文脈モデルに加える。したがって, 多品詞語や複合語, 並列構造などにより複数の木構造を生成する構造的曖昧性を含む文に関しては, その情報を文脈情報に加えずに, この段階では複数の候補をそのまま保持しておき, ステップ (2) で曖昧性を解消した後, そこで選択された木構造の情報を文脈モデルに加える。

各文の情報を文脈モデルに加える際には, 構文解析結果から係り受けの情報を抽出し, 表1に示すように, 語ごとにまとめて蓄積する。

(2) 複数の木構造を生成する文に対する木構造の選択
ステップ (1) で構文解析された文のうち, 多品詞語や複合語, 並列構造などにより複数の木構造が生成された文について, 文脈情報を用いて一つの木構造を選択し, その選択された木構造の情報を文脈モデルに加える。

その際, 対象とする文から解釈可能な各解析構造に対して,

- 文脈内に同じ語と同じ接続関係に係る語が存在する割合
- 文脈内に同じ品詞の同じ語が存在する割合
- 隣接文との文型の一致度

の各要素から文脈モデルとの適合度を数値化し, その値の最も高い構造を選択する。文脈情報との適合度のみでは一つの木構造に絞り切れない場合, 残った木構造に対し, 従来の手法を用いて各解析構造の文法的妥当性を評価し, 評価値の最も高い構造を選択する。こ

☆1 曖昧な複数の候補に対して異なる解釈を支持する情報が複数の文に存在する場合に, 文の位置情報を利用して, より近い文の情報を優先させることができる。例えば, 文1が解釈aを支持し, 文2が解釈bを支持する場合, 対象の文により近い文の支持する解釈を優先させる。ただし, 本論文の実験では, 文脈の範囲を限定する以外には位置情報を使用していない。

☆2 関係を表現するための記述として, 格関係名や接続している前置詞の表層語の他に, to 不定詞句や過去分詞句, 関係節として係り受けを結ぶことを表現するための記述を用いる。これによって, 前置詞句以外のタイプの係り受けの曖昧性も前置詞句と同様に扱うことができる。

☆3 後述の実験では構文解析結果を得るために, ESG パーザ⁶⁾およびPEG パーザ⁴⁾の2種類の構文解析器を使用した。文脈モデル構築にあたって構文解析器の違いによる影響は受けなかった。

☆4 外部知識を利用した曖昧性解消の情報が存在しない場合, デフォルト値として, 確定した係り先には選好度10, 曖昧な係り先には選好度3が与えられている。

表1 “define”の係り受けに関する文脈情報
Table 1 Example of discourse information on “define”.

品詞	接続関係	係ってくる語 (選好度)
代名詞	<i>SUBJ</i>	you (30)
	<i>OBJ</i>	it (10)
名詞	<i>OBJ</i>	application (10) network application (30) public application (10) additional server (20)
	<i>as</i>	application (10)
	<i>in</i>	domain (20)
	<i>to</i>	domain controller (3)
動詞	<i>PREINF</i>	use (3)
	<i>after</i>	install (10)
	<i>before</i>	start (10)
	<i>if</i>	install (10)
	<i>so that</i>	use (10)
品詞	接続関係	係り先の語 (選好度)
名詞	<i>PASTPART</i>	alias (20)
	<i>PREINF</i>	procedure (3)
動詞	<i>PREINF</i>	follow (3)
	<i>after</i>	use (10)

のようにして選択された木構造から、係り受け情報を抽出し、前ステップと同様に文脈モデルへ加える。

(3) 係り受けの曖昧性の解消

ステップ(1)で文脈内のすべての文が構文解析された後、ステップ(2)を通じてすべての文に対し唯一の構文解析木が選択される。このステップでは、各文の構文解析木に含まれている係り受けの曖昧性を解消していく。

係り受けの曖昧な句が存在すると、その係り先の各候補に関して同じ係り受け関係が文脈内に存在しないかどうかを調べ、同じ係り受け関係が存在する場合には、その頻度に重みを掛けた値を、その係り受けに対する選好度として与える。その結果与えられた選好度の最も大きな係り先を選択することで曖昧性を解消する。例えば、(3.1)(3.2)(3.3)の3文のみからなる文脈で、(3.1)や(3.3)における“on the MTO console”の係り先を決定する場合を考えると、名詞“console”が“on”を介して動詞“display”に係るボタンが、文脈内に3例^{*}存在し、確定的なボタンの重みを10、不確定なボタンの重みを3とすると、選好度16が係り先候補の動詞“display”に与えられる。その結果、係り先候補としての選好度が最も高くなるので、動詞“display”が係り先として選択される。

また、全く同じ係り受け関係のみに限らず、係る側か係られる側のどちらかを同義語あるいは類義語で置き換えたボタンも調べ、同じボタンの場合よりも小さい重みを掛けた選好度を与える。すなわち、係り受けの曖昧な句を P_{amb} とし、その係り先候補を P_{cand-i}

をすると、 P_{amb} が P_{cand-i} に係っているボタンだけでなく、 P_{amb} の同義語や類義語が P_{cand-i} に係っているボタン、あるいは P_{amb} が P_{cand-i} の同義語や類義語に係っているボタンもないかどうか調べる。その際、電子化された The New Collins Thesaurus¹⁶⁾ において同義語として記載されている語、および、文脈内で and や or により並列に扱われている語を、各々同義語および類義語^{☆☆}として扱う。

4.3 実験結果

4.3.1 文脈範囲と文脈情報のカバー率

まず、係り受けの曖昧性の解消において、文脈から情報を得られる割合が文脈の大きさ(文の数)によりどう変化するかを調べた。文脈の大きさを10文から791文に変化させた時に、文中で係り先の曖昧な句^{☆☆☆}に関して、文脈から情報を得られる割合が変化した様子を図2に示す。図中、横軸で示されるのは文脈の大きさで、単位は文数である。

(A)は、係り先の曖昧な句あるいは係り先候補の句のヘッドとなる語と同じ語が文脈内の他の文に存在する割合、すなわち文脈中で同じ語が繰り返される割合を示す。例えば、 $Verb_1$ $Noun_1$ $prep$ $Noun_2$ というボタンの曖昧性解消において文脈を参照する際に、文脈の範囲を80文以上にすれば、その範囲内の他の文で $Verb_1$, $Noun_1$ または $Noun_2$ が見出される確率が80%を越え、200文以上にすれば85%を越えると

* (3.2)の確定的なボタンと(3.1)(3.3)の不確定なボタン。

☆☆ 例えば、“In the MVS environment, the product runs under TSO/E, CICS, and IMS.”という文で“TSO/E,” “CICS,” および “IMS” が並列に扱われているので、この文を含む文脈内では、この3語を類義語として扱う。

☆☆☆ 前置詞句のみでなく、to不定詞句、分詞句、関係節も含む。

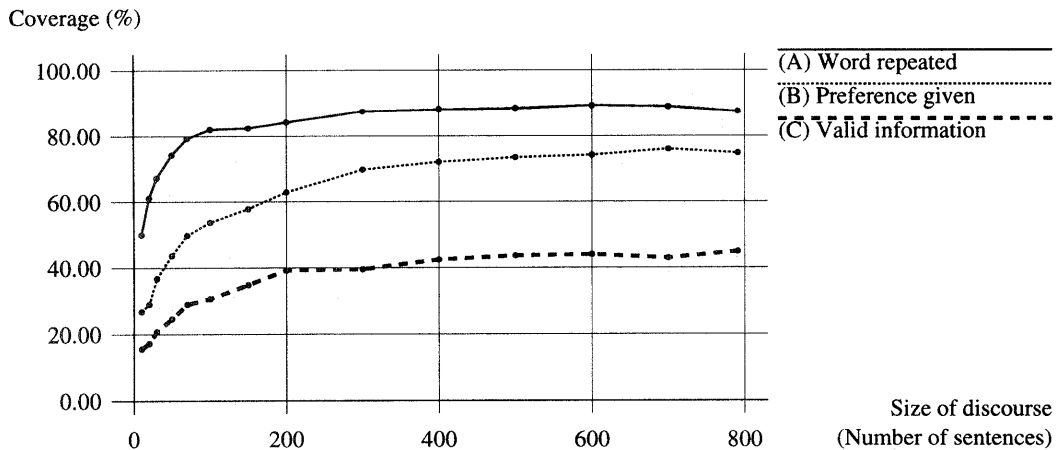


図2 文脈の大きさと文脈情報のカバー率の関係

Fig. 2 Relation between the coverage of discourse information and the size of the discourse.

いう結果が得られた。

(B) は、曖昧性解消の際に、文脈内の他の文の係り受け情報から選好度が得られる割合を示す。すなわち、係る側の句と係り先候補のうちのいずれかの句との係り受けパターンが他の文にも存在する割合を示している[☆]。文脈範囲が300文を越えれば、70%程度の確率で同じ係り受けパターンが文脈内で繰り返されているという結果が得られた。しかし、同じ係り受けパターンが文脈内に存在しても、全く同じ曖昧性を持つ係り受けのパターンが繰り返されているのでは、曖昧性解消に貢献しない。例えば(3.3)と全く同じ文が存在しても、“on the MTO console”の係り先候補の“display”と“CICS”の両方の選好度が増加するので、候補間の優位性に差が生じない。そこで、同じ係り受けのパターンが文脈内で繰り返され、かつ、それが(3.1)に対する(3.2)や(3.3)のように、曖昧性解消に貢献している割合を示したのが(C)である。

(C)で示されるように、文脈の範囲を200文程度に設定すれば、その中に含まれる係り先の曖昧な句の約40%を文脈内の情報で処理できるという結果が得られた。

4.3.2 前置詞句の係り先に関する曖昧性解消の成功率

次に、係り先の曖昧な句全体の約2/3^{☆☆}を占める前置詞句の曖昧性を解消する実験を行った。

ある計算機マニュアルの一つの章(218文)において、係り先の曖昧な前置詞句が109句存在し、最も近い係り先を選択するという単純な規則で係り先を決定

した場合の成功率は69.7%(76句)であった。これらの句を本手法で処理した結果、82.6%(90句)において正解が得られた。全体(109句)のうち、事例ベース中に蓄えられた外部知識が適用された^{☆☆☆}のは22句(20.2%)であり、そのうちの19句では事例ベース中の知識が正解に寄与したが、残る3句においては誤った係り先に高い選好度を与えていた。すなわち、事例ベース中に適用可能な知識が存在しても、必ずしも正解に結び付くとは限らないという結果が得られた。この文章で文脈情報に基づいて選好度を決定したのは31句(28.4%)であったが、そのうちの6句では誤った係り先を選択していた。この6句を調べてみると、そのうち3句においては、述語が省略された文の情報を参照していた。具体例をあげると、

(4.1) You can sign off your AS/400 session in one of the following ways:

に続く箇条書中の文、

(4.2) Using option 90 (Sign off) from the AS/400 Main Menu

において、“from the AS/400 Main Menu”の係り先を“Using”と“option 90”の二つの候補から選択する際に、

(4.3) Select one of the following options:

に続く箇条書中の文、

(4.4) Option 90 (Sign off) from the AS/400 Main Menu.

を参照し、(4.4)の文において“from the AS/400 Main Menu”の確定的な係り先として解析された“option 90”を(4.2)における“from the AS/400 Main Menu”

[☆] 同じ語のみでなく、同義語で置き換えられる場合も含めた。

^{☆☆} 我々の調査では、計算機マニュアルで67.4%であった。

^{☆☆☆} ここでは浦本¹⁴⁾の手法を用いた。

の係り先として優先していた。このように、省略の存在により構文解析で誤った結果が得られた場合に、その情報が文脈内の他の文の係り受けに悪影響を及ぼすという問題が見られた。残る3句においては、実際には選択されない離れた係り先候補に関する情報を参照していたもので、これらに関しては、曖昧性を含む情報を利用する際に、係り受けを結ぶ語の間の距離を考慮し、このようなノイズ的情報の影響を抑える工夫が必要である。

また、文脈情報を参照する際、前方の文における情報と後方の文における情報のどちらをより多く利用しているか調べたところ、前方19件、後方20件と、ほとんど差がなく、文間距離もバラツキの大きいことがわかった(表2)。

4.3.3 多品詞語や複合語、並列構造などにより生成される複数の木構造の選択

さらに、多品詞語や複合語、並列構造などの存在により生成される複数の木構造から、本手法を用いて文脈情報との適合度の高い構文木を選択した結果を示す。実験対象として計算機マニュアル中の連続した736文を用い、入力文章を400文と336文の二つの文脈に分割して処理した。全736文中、複数の木構造が生成された文は150文(全体の20.4%)存在したが、そのうち、複数の候補が、

- 微妙に異なっているが、ほとんど差がない
- すべて解析エラーで、どの候補を選択しても無意味

と判断された24文を抜くと、選択処理が必要で、かつ最適解の判断が可能な文は126文となった。この126文において、複数の解析候補から正しい構造を選択する上で、文脈適合度の各要素がどの程度有効かを調べた。その結果を表3に示す。最も有効性が高かったのは、同じ係り受けパターンが文脈内の他の文に存在する割合の大きさであった。126文全体の42.9%にあたる54文において、その割合の大きさに差が認められ、そのうち88.9%にあたる48文で、最適解に最高値が与えられていた。この実験から、本論文の曖昧性解消手法で仮定した文脈内の一貫性、すなわち、文脈内で、

- 同じ語は同じような語と係り受けを結ぶ
- 隣合う文は同じ文型を取る

という傾向が、実験対象の計算機マニュアルでは成立していると見なせる結果が得られた。

次に、本手法により、構文解析の精度が実際にどの程度向上するかを上記の調査とは別の文章で調べた。ここではESGパーザ⁶⁾を用い、パーザの出力する解

析結果において複数の候補が存在する場合、以下の3ステップで一つの構造を決定した。

- (I) 同じ語と同じ接続関係の係り受けパターンが文脈内に存在する割合の最も大きい解析結果を選択
- (II) ステップ(I)の選択結果が複数存在する場合、隣接文との文型の一致度の最も大きい解析結果を選択
- (III) ステップ(II)の選択結果が複数存在する場合、ESGパーザの出力した文法的な妥当性の評価値の最も大きい解析結果を選択

この手法を用いて、計算機マニュアル中の1章を構成する全244文を一つの文脈とし、その全文をESGパーザで解析したところ、110文において複数の解析候補が出力された。そのうち、文法的な妥当性に基づくESGパーザの評価値のみで正しい候補を選択できたのは82文(精度74.5%)であったのに対し、本手法を用いることで正しい候補を選択できる文が98文(精度89.1%)になった。この結果、244文全体の解析精度という観点からは、ESGパーザのみで解を選択した場合の精度が71.3%であったのに対し、本手法を用いて複数の解析候補からの選択を行うことで、精度が77.9%へ向上した。すなわち、複数文の選択に関しては74.5%から89.1%へ、文章全体の解析に関しては71.3%から77.9%へ精度が向上するという結果が得られた。全244文中で、正しい解析結果を得られなかった文の内訳を表4に示す。

5. おわりに

文脈情報を利用した構造的曖昧性の解消手法を提案し、実験結果を通してその有効性を示した。本手法による曖昧性解消の精度は、文脈内に存在する情報に依存しており、一貫したテーマに関して同じような内容を繰り返して記述している計算機マニュアルのような文章では、曖昧性解消の際に参照できる文脈情報が豊富なため、本手法の効果が大きく、実験でも有効な結果が得られている。また、本手法は文脈処理用の特別な知識を前提とすることなく、単純な処理で解析精度を向上させられるため、頑健で実用性が高い。文脈全体の構文解析結果を保持しておく点で、従来の手法よりも大きな記憶容量を必要とするが、ハードウェアの進歩により、利用可能な主記憶容量は十分に大きくなっており、システムの実用性を低下させる要因とはならない。また、文脈内の情報を参照する処理も、従来の手法で知識ベースを参照する処理と同等の計算量であり、大きな負荷となることはない。

しかし一方、文脈内に存在する情報に依存している

表2 参照情報を含む文までの文間距離の分布

Table 2 Distribution of information relative to a sentence that contains information for modification preference.

前方	文間距離 (文数)	1 ~ 5	6 ~ 10	11 ~ 20	21 ~ 30	31 ~ 40	41 ~ 50	51 ~ 60
	件数	5	2	3	0	5	4	0
後方	文間距離 (文数)	1 ~ 5	6 ~ 10	11 ~ 20	21 ~ 30	31 ~ 40	41 ~ 50	51 ~ 60
	件数	6	0	3	1	6	3	1

表3 複数構文構造の選択における文脈情報との適合度の各要素の有効性

Table 3 Effects of each factor on multiple parse selection.

適合度の要素	候補間で差が出る	最適解が最高値を取る
文脈内に同じ語と同じ接続関係に係る語が存在する割合	42.9% (54 文/126 文)	88.9% (48 文/54 文)
文脈内に同じ品詞の同じ語が存在する割合	44.4% (56 文/126 文)	66.1% (37 文/56 文)
隣接文との文型の一致度	49.2% (62 文/126 文)	77.4% (48 文/62 文)

表4 文脈情報を用いた複数構文構造の選択による構文解析精度の向上

Table 4 Improvement of multiple parse selection based on discourse information.

構文解析において正しい解析結果を得られなかった文の内訳	(a) 解析結果が全く得られなかった	9 文
	(b) 一つの構文構造に達しなかった	22 文
	(c) 唯一の解析結果が得られたが誤った解析結果であった	11 文
	(d) ESG パーザの評価値のみでは正しい解析結果を選択できなかった	28 文
	(e) 本手法を用いても正しい解析結果を選択できなかった	12 文
本手法を用いない場合の解析精度 (244-(a)-(b)-(c)-(d))/244		71.3%
本手法を用いた場合の解析精度 (244-(a)-(b)-(c)-(e))/244		77.9%

ことから、本手法のみで完全な解析精度を実現できるわけではなく、あらかじめ構築した知識を用いる従来の手法との併用が望ましい。その場合、従来の手法であらかじめ構築しておく知識を長期記憶、本手法で用いる文脈内の情報を短期記憶とみなすことができる。長期記憶の知識は記憶装置の許す範囲でいくらかでも蓄積していくことが可能であるが、文章はテーマや書き手による多様性が大きく、現実的には、あらかじめ用意した知識の適用範囲に限界がある。それに対し、短期記憶を導入し、文脈内の一貫性を確立する処理を適用することで、知識の不足を補うと共に、文脈依存性の問題にも対応することが可能になる。知識として蓄えられた長期記憶と異なり、短期記憶としての文脈情報は、それ自体が曖昧性を内包している。しかし、すべてが曖昧な情報ではない上に、複数の曖昧な情報が相互に補い合うことで曖昧性を解消できるパターンも存在し、各々の文を別々に処理していたのでは得られない情報を獲得することもできる。さらに、文脈情報を利用する処理は、文脈に適した解釈を優先させられる反面、省略などの存在により誤って解析した情報も文脈内に伝播してしまう危険性も備えている。しかし、これまでの実験においては、その割合は有効な情報の伝播する割合を十分に下回っており、また、省略を含んでいると判断される文の情報は文脈情報に加えないなどの対応も考えられる。

本論文では英文を対象としたが、係り受けの曖昧性解消における、文脈内から抽出した係り受け情報の有効性は、木下ら⁵⁾によって日本語文の解析でも成り立つことが示されている。文脈内の情報は、同じ書き手が同じテーマに関して書いた文で構成されており、文どうしの類似性が高いところから、文の曖昧性を解消する上で、最も適用性の高い有効な知識源として捉えることができる。その観点からも、係り受けや、語義、並列されている語など、文脈内から抽出できる情報を最大限に利用する意義は大きい。

謝辞 本研究に関して重要な示唆を与えて下さった自然言語処理グループの堤泰治郎、諸橋正幸、武田浩一、丸山宏、野美山浩、渡辺日出雄、荻野紫穂、浦本直彦の諸氏に心から感謝いたします。

参考文献

- 1) Gale, W., Church, K. and Yarowsky, D.: One Sense Per Discourse, *the 4th DARPA Speech and Natural Language Workshop* (1992).
- 2) Hindle, D. and Rooth, M.: Structural Ambiguity and Lexical Relations, *Computational Linguistics*, Vol.19, No.1, pp.103-120 (1993).
- 3) Isahara, H. and Ishizaki, S.: Context Analysis System for Japanese Text, *COLING-86*, pp.244-246 (1986).
- 4) Jensen, K.: PEG: The PLNLP English

- Grammar, *Natural Language Processing: The PLNLP Approach*, Jensen, K., Heidorn, G. and Richardson, S. (eds.), Kluwer Academic Publishers, Boston, Mass. (1992).
- 5) Kinoshita, S., Shimazu, M. and Hirakawa, H.: Better Translation with Knowledge Extracted from Source Text, *TMI-93*, pp.240-251 (1993).
 - 6) McCord, M.: The Slot Grammar System, *IBM Research Report*, RC17313 (1991).
 - 7) Nagao, K.: Semantic Interpretation Based on the Multi-World Model, *IJCAI-89*, pp.1467-1473 (1989).
 - 8) Nagao, K.: Dependency Analyzer: A Knowledge-Based Approach to Structural Disambiguation, *COLING-90*, pp.282-287 (1990).
 - 9) Nasukawa, T.: Discourse Constraint in Computer Manuals, *TMI-93*, pp.183-194 (1993).
 - 10) 那須川：文脈制約を利用した曖昧性解消，第7回人工知能学会全国大会，pp.425-428 (1993).
 - 11) Sumita, E., Furuse, O. and Iida, H.: An Example-Based Disambiguation of Prepositional Phrase Attachment, *TMI-93*, pp.80-91 (1993).
 - 12) Takeda, K., Uramoto, N., Nasukawa, T. and Tsutsumi, T.: Shalt2-A Symmetric Machine Translation System with Conceptual Transfer, *COLING-92*, pp.1034-1038 (1992).
 - 13) 浦本：制約と事例による優先度を組み合わせた英文の多義性の解消，情報処理学会自然言語処理研究報告 (92-NL-90) (1992).
 - 14) Uramoto, N.: A Best-Match Algorithm for Broad-Coverage Example-Based Disambiguation, *COLING-94*, pp.717-721 (1994).
 - 15) Utsuro, T., Matsumoto, Y. and Nagao, M.: Lexical Knowledge Acquisition from Bilingual Corpora, *Proceedings of COLING-92*, pp.581-587 (1992).
 - 16) *The New Collins Thesaurus*, Collins Publishers, Glasgow (1984).

(平成7年1月5日受付)

(平成7年7月7日採録)

那須川哲哉 (正会員)



1987年早稲田大学理工学部電子通信学科卒業。1989年同大学院修士課程修了。同年、日本アイ・ビー・エム(株)入社。現在、同社東京基礎研究所において機械翻訳、電子図書館の研究開発に従事。1993年度人工知能学会全国大会優秀論文賞受賞。人工知能学会、言語処理学会、AAAI、ACL各会員。