

並列協調に基づく自然言語解析実験システム *Laputa*

目的

本研究の目的は並列推論マシンの利点を活用して自然言語処理のあらゆる処理(形態素解析, 構文解析, 意味解析など)を統合するためのソフトウェア技術の基礎を開発することである。

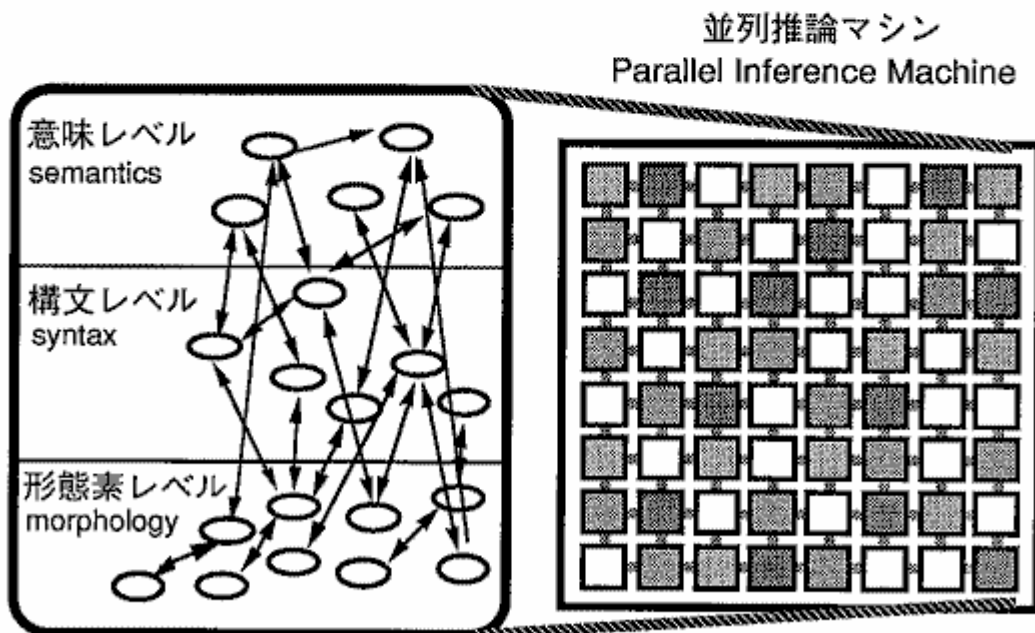
概要及び特徴

(1) 自動的な並列協調処理の実現

自然言語処理の全ての部分処理を共通の処理機構で処理する並列協調モデルを提案した。

(2) 性能

32台構成のマルチPSIで, 1台での処理と比較して13倍の速度向上率を達成することができた。



システム構成

研究の趣旨

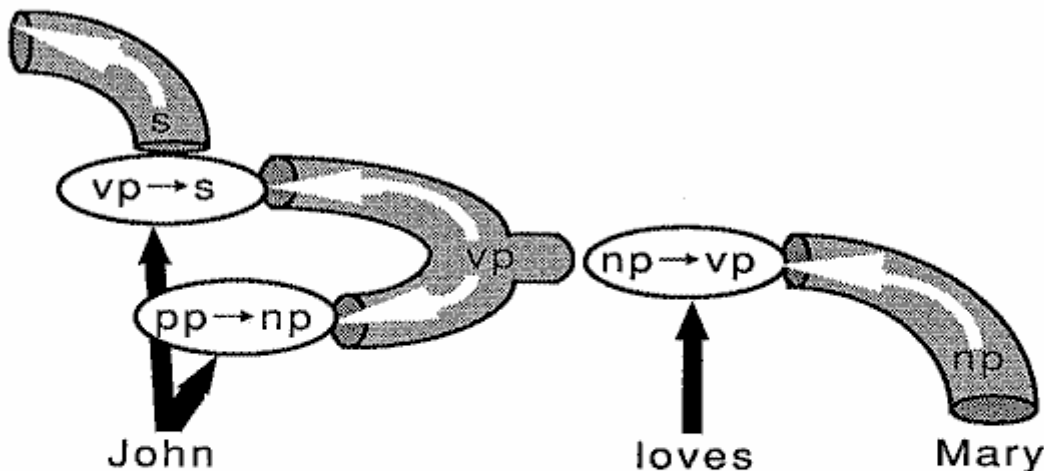
本研究の目的は、並列推論マシンの利点を活用することにより、自然言語処理のための自然な並列協調モデルを提案することである。

近年、自然言語処理の分野では、形態素解析、構文解析、意味解析などを含めたシステムの統合が提案されている。このような考え方の基盤として、人間の情報処理が情報の部分性あるいは不完全な情報に基づいて行われているという認識がある。このように統合された自然言語処理は、処理の方向性を捨象しているという意味で並列処理に適したものである。したがって、自然言語の統合と並列協調は自然な処理モデルといえることができる。

しかし、そのようなモデルを実現した例はほとんどない。なぜなら、効率的に並列協調を行うには全ての処理過程は相互に情報を交換する必要があるが、モジュール間での情報交換やその制御は非常に実現困難な問題だからである。このような問題に対する一つの解答は、解析の段階を区別なく同一の機構で処理するように処理の枠組みを抽象化してしまうことである。我々は処理の枠組みとしてレコード的型構造に関する型推論を採用した。

アプローチ

形態素解析や構文解析に関しては効率的なアルゴリズムが存在しており、統合的自然言語処理においても、実用的なシステムを構築するにはそういった知識を無視するわけにはいかない。



レイヤードストリーム法による構文解析

幸いなことに我々は型判断と既知の構文解析手法との間に関係を見いだすことができた。松本による並列構文解析システム PAX は「レイヤードストリーム法」という並列論理型言語による探索問題のための効率的な処理手法を用いて構文解

析を行うものである。PAX は基本的にチャート法と同等の構文解析アルゴリズムになっている。我々の型判断システムは PAX のプロセスと通信データの関係性を反転させたものになっている。また、我々のシステムと同等の構文解析システムが ICOT の瀧によって考案されている。

文法及び辞書記述の例 [主語と目的語の意味的關係の動的決定]

以下の文法規則例は、主語に依存していかに関数型の意味範疇の適切性が動的に変更されるかを示している。

$$\{np, [sem=Subj]\} < (\{vp, VP\} \rightarrow \{s, VP=[agent=Subj]\})$$

この規則は、型 np の上位の型が関数型 “vp → s” であることを示しており、あるオブジェクトが np として判断されたならば、それは vp という型を持つオブジェクトが適用されると結果として s という型のオブジェクトになるような関数でもあるということの意味している。この規則では vp が持つ全ての記述は s にマージされ、np の sem という素性の値は s の agent という素性の値と単一化される。

次にこの規則に対応する語彙項目を掲げる。

eats: {np, [sem=Obj]} → {vp, [agent=Ag: {animal, [eat_obj=Obj]}]}

john: {np, [sem={human, Id, [name='John']}]}

the_tiger: {np, [sem={tiger, Id, []}]}

この辞書ではオブジェクト eats は np → vp という型であり、さらに記述が付加されている。その記述に書かれているのは、agent という素性の値は型付き変数であり、その型は animal というソートであって、その eat_obj という素性の値が型 np の sem 素性の値と単一化されているということである。

意味範疇を定義する規則は次のものとする。

$$\begin{aligned} \{tiger, []\} &< \{animal, [eat_obj=E:\{animal, []\}]\} \\ \{human, []\} &< \{animal, [eat_obj=E:\{food, []\}]\} \end{aligned}$$

この規則の意味は、トラは動物を食べる動物であり、人間は食べ物を食べる動物であるということである。

この文法規則と辞書のもとでは john も the.tiger も意味範疇は動物であるが、(the.tiger, eats, john) : s という判断は成功するのに (john, eats, the.tiger) : s という判断は失敗する、その理由はこの意味範疇を統御する規則のもとではトラは食べ物として判断できないからである。

システムの概要

文法及び辞書の規模は次のとおりである。

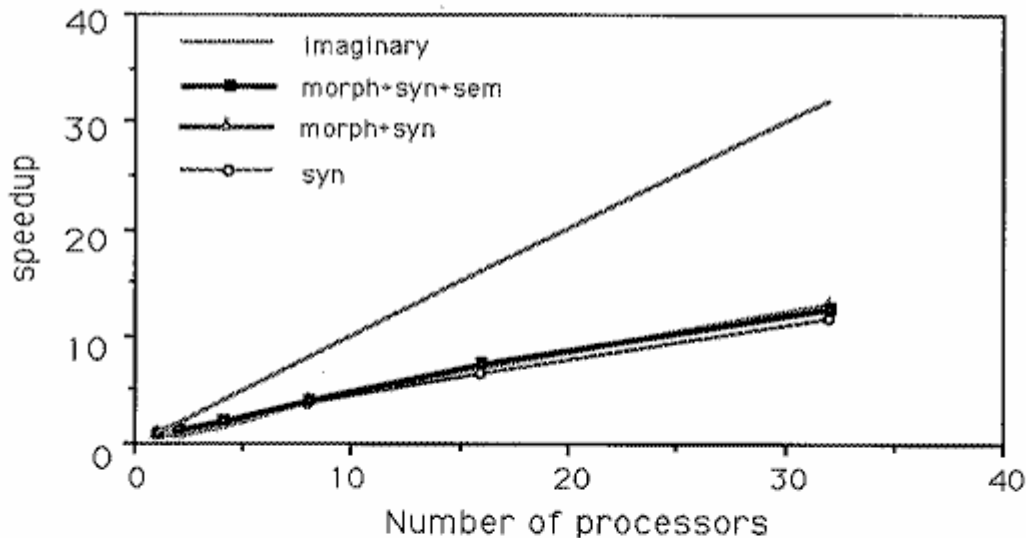
文法規則	651
単語数	14,613
形態素数	8,268
概念数	770

形態素解析文法及び構文解析文法については ICOT 第 6 研究室の佐野が開発した文法を利用した。概念体系については EDR の概念体系に準拠した。

結果

実験の結果、32 台のプロセッサを使用したときに約 13 倍の速度向上率を実現できた。

台数効果



デモ内容

1. 自動的並列協調の実演
2. 主語と目的語の意味関係が動的に決まることの実演
3. 形態素解析, 構文解析の並列協調処理の実演
4. 意味処理を加えた並列協調処理の実演
5. 台数効果を見る