決定性の中国語依存構造解析器の改良

鄭育昌　浅原正幸　松本裕治
奈良先端科学技術大学院大学　情報科学研究科
{yuchan-c, masayu-a, matsu}@is.naist.jp

1　はじめに

近年、意味解析、照応解析及び質問応答などの分野において、依存構造解析の結果を用いた研究が盛んである。英語 [Yamada, 2004; Nivre, 2004]と日本語 [Kudo, 2002]の依存解析についてはすでに有効なアプローチが提案されてきた。しかし、中国語には依存構造解析についての研究は少ない。Zhou [Zhou, 2000]とLai [Lai, 2001]はルールベースの依存構造解析を提案した。しかし、ルールベースの手法では人手で規則を構築する必要があり、専門的な言語知識が必須である。他の研究として、Ma [Ma, 2004]は小規模コーパスで統計ベースの実験を行った。しかし、Maの統計モデルは単語数 n の入力文に対して$O(n^3)$のアルゴリズムであり、解析が時間を掛かる。本稿では、$O(2n)$の上昇型決定性のアルゴリズム[Nivre, 2004]を用いて、中国語の依存構造解析器を実装し、その依存構造解析器に対して、いくつかの拡張を提案した。まず、機械学習分類器の手順及び素性選択の欠点を考慮し、大域素性と二段式分類アプローチを導入することで依存解析器を改良した。次に、磯崎ら[Isozaki, 2004]の研究に提案されたルート解析器の導入をさらに拡張して、依存構造の特性を考慮して文の分割による解析手法を導入し、正解率を向上させた。

2　上昇型決定性中国語依存構造解析器

2.1　先行研究

本稿で導入する上昇型決定性依存構造解析のアルゴリズムはNivre [Nivre, 2004b]の提案手法を基としている。Nivreのアルゴリズムは Yamada [Yamada, 2004]のアルゴリズムの問題点を改良した、新しいアルゴリズムである。Yamadaアルゴリズムにおいて、単語の依存関係を決定する動作に曖昧性があることが指摘されている[Cheng, 2004]。Nivreアルゴリズムはその曖昧性を解消できると予想されるが、英語の実験結果はNivreが提案した手法よりYamadaが提案した手法の方が正解率が良かった[Nivre, 2004]。これはアルゴリズムの差異ではなく、機械学習方法の差異によるものと考えられる。Yamadaの手法はSupport Vector Machines (SVMs)を使用しているが、Nivreの手法はMemory Based Learningを使用いる。同じ機械学習方法を使用して、両アルゴリズムを中国語に実装した結果[Cheng, 2004]によると、Nivreアルゴリズムと SVMsを用いる解析器がより優れた結果を得られることが分かった。

2.2　基本的な解析手法

本研究の基礎となる依存構造解析器は Nivre アルゴリズムによる。Nivre アルゴリズムは$\langle S, I, A\rangle$で状態を表示する。S と I はスタックで、S は解析済みのトークン列、 I は未解析のトークン列を格納する。 A は依存関係を記録するリストである。解析目標の単語と品詞からなるトークン列 W に対して、初期状態は $\langle nil, W, \phi\rangle$ である。アルゴリズムは常にスタック S の最後のトークンとスタック I の最初のトークンの間の依存関係を解析する。解析はスタック I が空になる時点で終了する。解析しているトークンペアの可能な関係及び関係定義後の動作は下記の４つである（図２参照）：

Right: 解析しているトークンペアの左トークンが右トークンに係る。この依存関係を A に記録して、左トークンをスタックから消す。

Left: 解析しているトークンペアの右トークンが左トークンに係る。この依存関係を A に記録して、右トークンをスタック I からスタック S に移動する。

Reduce: 左トークンがスタック S にあるいずれかのトークンに係る。かつスタック I には左トークンの修飾語がない場合、左トークンがスタックから消す。

Shift: Reduce の状況を満たさない場合、右トークンがスタック S に移動する。

上記の動作を機械学習分類器を用いて決定する。本稿では SVMs を使用して動作を決定する。機械学習に使用される素性は図１のように解析しているトークンペアの前後２トークンの情報、及びトークンの距離などである。

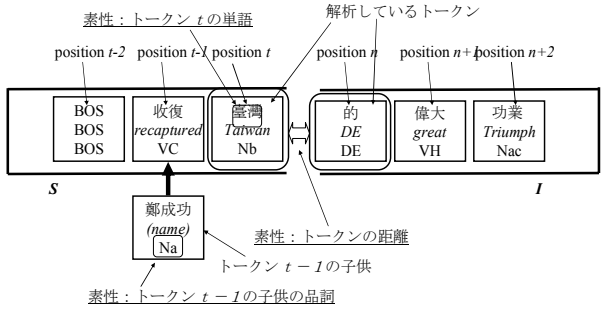

図 1:基本素性

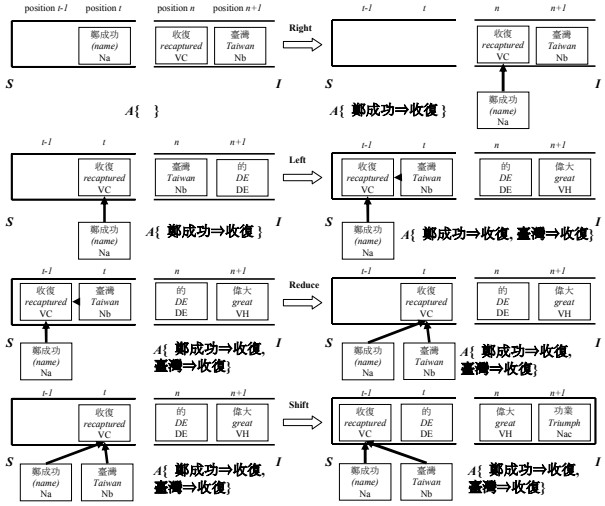

図 2：Nivre アルゴリズムの動作

3　提案手法

本稿では、基となる技術として、Nivre らが提案した上昇型の構文解析アルゴリズムと機械学習器 SVMs を用いる。Penn Chinese Treebank を用いた評価実験および誤り分析の結果、大域素性の必要性とルート同定の困難さがわかった。この問題点に対して、２つの拡張を提案する。１つ目の拡張は４択の分類問題を小問題に分割し、必要なお小問題にのみ大域素性を用いて分類する手法である。２つ目の拡張は、磯崎らが提案したルート解析器を用い、長い文を２つに分割して解析する手法である。

3.1　大域素性及び二段階分類

図１に示すように、素性として注目しているトークンペアの前後２トークンの情報を用いている。Nivre アルゴリズムの定義に従って、解析しているトークンの間には依存関係がない場合、動作 **Shift** と **Reduce** が可能である。しかし、**Reduce** が成り立つ条件：「スタック *I* には左トークンの修飾語（子供）がない」に対して、図１のような局所的な素性を用いてこの条件を正確に判断することは不可能である。なぜなら、図１の素性はスタック *I* の全トークンを含まないためである。この基本素性を以後局所素性と呼ぶ。

図３の点線で囲む素性は局所素性である。注目しているトークンペアに対して、文脈素性として修飾語「出發（出発）」が分類器に与えられていない。このため、トークンペア間に係り受け関係がないことがわかっても、文脈素性としてスタック *I* の最後の形態素「出發」を含んでいないため、左トークン「我（わたし）」に係るトークンがスタック *I* にあるかどうか判定することは困難である。動作の判定は誤って **Reduce** になる。図３中の実線で囲まれた素性を使えばトークン「我」に係るトークンの有無を判定でき、動作 **Shift** と **Reduce** の区別を正確に解析できる。動作の判定は正しい動作 **Shift** になる。この図３中で実線で囲まれた素性を大域素性と呼ぶ。この大域素性は動作 **Right** と **left** の解析にはあまり有用ではないだろう。そこで、有効に局所と大域素性を活用するために、二段式の解析を導入する。図３に示す解析手法は二段式解析である。まず、分類器は局所（基本）素性を選択して解析する。ここは両トークンの間に依存関係か否かを判定する。出力が動作 **Shift** と **Reduce** の場合、即ち両トークンは依存関係がない。この場合にのみ解析器は大域素性を用いて、**Shift** と **Reduce** を識別するために再解析する。

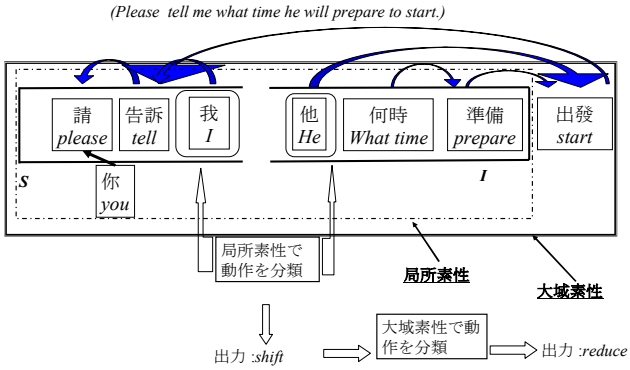

図 3：大域素性と二段式解析

3.2　ルート解析器の導入

磯崎ら[Isozaki, 2004]の提案手法はルート解析器を導入して、解析したルート情報を依存関係解析の素性として導入する。しかし、中国語において、ルート情報を単なる素性として使用する手法は依存構造解析の正解率を向上できなかった。本節で

は、解析したルートを用いて、文を分割して解析を行う手法を検討することである。

依存構造の原則によると、構造木の依存関係はルートを越えることができない。従って、ルートで構造木を分割して、２つの部分依存木とみなすことが可能である。図４にルート解析結果を用いて文を分割する手法を示す。図４の上半部の入力文に対して、まず、ルート解析器を用いて入力文のルート「與（と）」を決定する。入力文がルートを含んだ２つの部分木になる（左と右部分木）。依存構造解析は各部分木毎に行う。

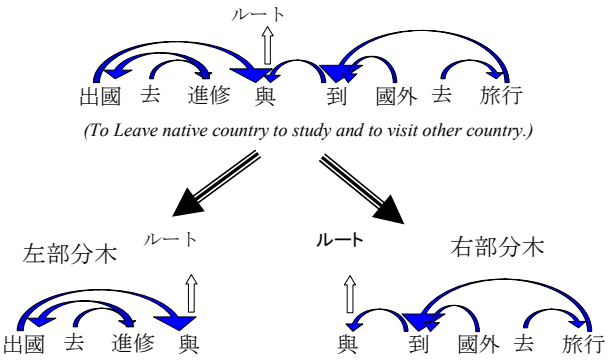

図 4:ルートで文を分断

4 実験環境

4.1 コーパス

本研究は機械学習に基づく手法であり、訓練事例として依存関係のタグ付きコーパスが必要である。Penn Chinese Treebank は句構造でタグづけされたコーパスのため、各フレーズの中心語（head）を決定する規則を構築する必要がある。我々は[Collins, 1999]と[Chen, 1995]を参考して、ヘッドルールを構築した。ヘッドルールに基づき Penn Chinese Treebank の句構造木を依存構造木に整形する。機械学習に使われる訓練事例は 140,200 語（10,000 文）、テストデータは 11,045 語（1122 文）である。Penn Chinese treebank の文章は新聞記事であるため、長文が多く、長距離依存構造を多く含む。

4.2 ルート解析器

本節に使用するルート解析器は磯崎ら[Isozaki, 2004]の論文に基づいて実装した解析器である。ルート解析器は機械学習（SVMs）で実装し、入力トークン列の各トークンにルートであるかどうかを解析してタグを付ける。図５はルート解析器に用いる素性を表示する。解析しているトークンの前後に単語の情報及びそれ以外の単語列情報などを素性として使われている。ルート解析器の訓練事例は前節の方法で整形されるコーパスの訓練事例である。本実験で用いたルート解析器の正解率（Recall）は 90.25%であり、磯崎らの英語における実験結果（95%以上）より低い。これは、適用した言語とコーパスの大きさが異なるためであると考える。

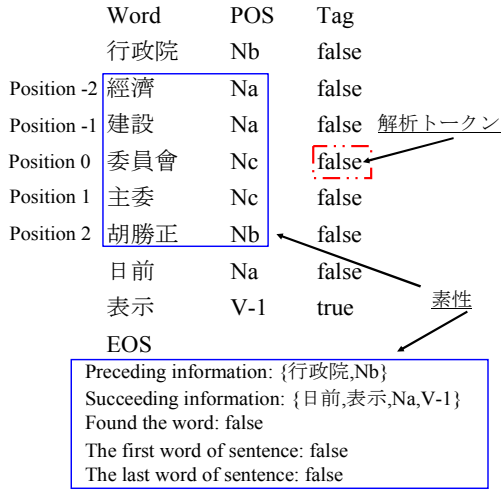

図 5：ルート解析の素性

5 評価実験

5.1 実験設定

本章は３章で提案した手法（大域素性の導入及びルート解析器の使用）の評価実験を示す。[Cheng, 2004]でNivreアルゴリズムをSVMsで実装する依存構造解析器をベースラインとし、提案する拡張がどのくらい有効であるか調査することを目的とする。ベースライン、大域素性と二段式解析を導入する手法、ルート解析器を導入する手法、両拡張手法を併用する手法を比較する。本研究の機械学習器はLibSVM []を使用して、ペアワイズ法 []で学習するである。XEON 2.4GHzのCPU と 4.0GB メモリーの環境で最大学習時間は32時間である。

5.2 実験結果

実験結果を表１に示す。評価方法は[Yamada, 2004]と同様に以下の３つの指標による：

Dependency Accuracy:

$$(Dep.\ Acc.) = \frac{\text{正しい単語の依存関係の数}}{\text{全ての依存関係の数}}$$

Root Accuracy:

$$(Root\ .\ Acc\ .) = \frac{\text{正しい文のルートの数}}{\text{文の数}}$$

Sentence Accuracy:

$$(Sent.\ Acc.) = \frac{\text{正しい依存構造木の数}}{\text{文の数}}$$

表１によると、両改良手法はともに依存構造解析器の正解率を多少向上させることがわかった。精度向上は小さいが、マクネマー検定により提案手法が優位であることをわかった。ベースラインと比べ、「ベースライン+大域素性」のp-valueは0.022<0.05、「ベースライン＋ルート解析器」のp-valueは0.007<0.05であった。両改良手法を併用すると依存構造正解率はさらに向上することがわかった。

	Dep. Acc.	Root Acc.	Sent. Acc.
ベースライン (Nivre with SVMs)	87.64	87.06	63.66
ベースライン+大域素性	87.82	87.10	63.79
ベースライン＋ルート解析器	87.93	**90.83**	**65.23**
ベースライン＋大域素性＋ルート解析器	**88.00**	**90.83**	65.20

表 1：評価実験の結果

6　討論

大域素性と二段式分類の手法は大幅には正解率を向上できなかった。その原因は各動作の事例の分布に偏りがあるためと考えられる。（**Right:** 44.7%; **Left:** 14.5%; **Shift:** 33.9%; **Reduce:** 6.8%）動作 **Reduce** の割合は最も少なく、大域素性と二段式分類が有効に効いている事例が少ない。大域素性を導入する目的は局所素性が考慮できない長距離の依存関係を配慮するためであった。長距離の依存関係は即ち解析している単語との距離が３以上の関係である。図２の「我（わたし）」と「出發（出発）」の関係は長距離の依存関係である。距離が３以上の長距離の依存関係は中国語にある割合が少ないため、この現象は大域素性と二段式分類の手法が大きく正解率を向上できない理由の１つと考えられるである。

ルート解析の結果で文を分割する手法について、大域素性の導入により正解率を向上できることがわかった。これは、文の分割により長い文を２つの短い部分に分割し、長距離の依存構造を含む組み合わせ爆発を未然に防ぐことができたことによる。しかし、この拡張による正解率の向上はルート解析器の正解率に大きく依存する。Penn Chinese Treebank の文が長くて複雑であり、ルートを先に解析することは困難である。従って、本稿のルート解析器の正解率は 90.25%であり、さらに向上することが必要となる。今後、ルート解析器の正解率をより向上させる手法を検討する。

7　おわりに

本稿では機械学習に基づく決定性の中国語依存構造解析器の改良手法について述べた。基となる技術として、Nivre らが提案した上昇型の構文解析アルゴリズムと機械学習器 SVMs を用いた。Penn Chinese Treebank を用いた評価実験を行い、誤り分析の結果、大域素性の必要性とルートの同定の困難さが判明した。１つ目の問題に対処するため、構造同定の手順を分割し、大域素性を用いたモデルと局所素性を用いたモデルを直列接続する手法を提案した。２つ目の問題に対処するため、磯崎らが提案したルート解析器を用い、長い文を２つに分割して解析する手法を提案した。最後に提案した拡張の評価実験を行い、有効性を検証した。

参考文献

Keh-Jiann Chen, Chu-Ren Huang, 1995. *訊息為本的格位語法與其剖析方法*. Technical report no. 95-03.

Yuchnag Cheng, Masayuki Asahara and Yuji Matsumoto, 2004. *機械学習に基づく決定性の中国語依存構造解析器*.自然言語処理研究会, 2004-NL-163, pp.91-98

Michael Collins, 1999. *Head-Driven Statistical Models for Natural Language Parsing*. PhD Dissertation, University of Pennsylvania

Hideki Isozaki, Hideto Kazawa, Tsutomu Hirao, 2004. *A Deterministic Word Dependency Analyzer Enhanced With Preference Learning*, COLING-2004

Taku Kudo and Yuji Matsumoto. 2002, *Japanese Dependency Analyisis using Cascaded Chunking,* CONLL 2002 in TAIPEI

Tom B. Y. Lai, C. N. Huang, Ming Zhou, Jiangbo Miao, T. K. C. Siu, 2001. *Span-based Statistical Dependency Parsing of Chinese*. NLPRS 2001: pages 677-68

Chih Jen Lin, 2001.*A practical guide to support vector classification*,http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Ma Jinshan, Zhang yu, Liu ting, and Li sheng, 2004. *A Statistical Dependency Parser of Chinese under Small Training Data*. IJCNLP 2004

Joakim Nivre, 2004. *Incrementality in Deterministic Dependency Parsing*. ACL-2004.

Hiroyasu Yamada and Yuji Matsumoto, 2004. *Support Vector Machine を用いた決定性上昇型依存構造解析* 情報処理学会論文誌, Vol.45, No.10, pp.2416-2427

Ming Zhou, 2000. *A block-based robust dependency parser for unrestricted Chinese text,* The second Chinese Language Processing Workshop attached to ACL 2000.