SVMを用いた不要な曖昧性の除去による構文解析高速化の検討

定政 邦彦, 安藤 真一, 土井 伸一
日本電気株式会社メディア情報研究所
k-sadamasa@az.jp.nec.com, s-ando@cw.jp.nec.com, s-doi@ah.jp.nec.com

1 はじめに

多くのルールベースの構文解析器は、自然言語の単語や句が意味の曖昧性を持つことを考慮して網羅的な解析が行えるよう設計されている。しかしながら、機械翻訳などの構文解析の実利用シーンでは、解析に用いる辞書や文法を固定し、事例データやヒューリスティックを利用したスコアリング関数を定めることで、構文解析器が出力する解析候補に優先順位をつけ、そのうちの第一位解のみを出力・利用することが多い。第一位解のみ利用できればよい用途では、第一位解以外の解釈を出力するための解析は処理の無駄であり、解析速度を低下させる原因になっていた。

これら不必要な曖昧性を判別する統計モデルを予め作成しておき、実行時に判別モデルに基づいてそれら不必要な曖昧性を早期に判別・削除することができれば構文解析処理を高速化することが可能となる。判別モデルとしては、構文解析によって解消された曖昧性を正解として学習し最尤推定するという方法も考えられるが、bigramやtrigramといった限られた情報では解決できない曖昧性に関して誤った曖昧性解消が生じるという問題点がある。

本稿では、明らかに不要と思われる曖昧性のみを判別し削除することで構文解析結果をできるだけ変えることなく処理を高速化する手法を提案する。判別モデルは、構文解析によって解消される曖昧性の情報をクラス、各曖昧性の周囲の文脈を素性としてSVM[6]で学習するが、学習データ中の対立する事例に修正を加えることで誤った曖昧性解消を低減したモデルを生成する。本稿では特に、形態素解析結果の不必要な曖昧性を判別・削減するモデルの構築について検討した。

本稿の構成は以下の通りである。２章では、曖昧性削減モデルの基本アイディアについて述べ、３章で詳細にモデルの構築方法を述べる。続いて４章ではSVMを学習器に使う際の問題点と対処を述べる。５章では各種条件下で実験を行って本モデルの性能を提示し、最後に６章で本稿をまとめる。

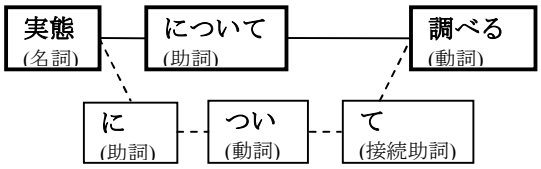

図 1：不要な曖昧性の例

2 不必要な曖昧性とは

既に述べたように、構文解析器が実アプリケーションで利用される際には、特定のスコアリング関数に基づいた第一位の解析結果しか用いられないことが多い。それにも拘らず、網羅的な解釈を出力可能に設計されている構文解析器、特にルールベースの構文解析器においては、第一位解を求めるのに必要のない曖昧性の計算まで不必要に行う可能性がある。例えば「について」という表現には、図１に示すように、解釈の可能性としては「長単位助詞」と「助詞＋動詞＋接続助詞」の２解釈が存在し、構文解析器も２解釈の解析を行う。しかし、実際にはごく一部の文脈(「“席”について」等共起が強く働く場合)を除いては長単位助詞の解釈しか選択されないようなスコアリングがされているとすると、構文解析器は多くの場合に、全く無駄に「助詞＋動詞＋接続助詞」の解釈を計算していることになる。

ここで、現在用いられているスコアリングにおいては「について」の長単位助詞の曖昧性しか残り得ないことが直前直後の文脈から予め判断できれば、「助詞＋動詞＋接続助詞」という不要な曖昧性に関する計算をスキップでき、構文解析処理を高速化できる。本稿では、このように、削除しても構文解析結果を変化させない曖昧性を周囲の文脈から判別するモデルを予め構築し、解析時に利用することで構文解析処理を高速化する手法を提案する。

本手法を適用可能な曖昧性の種類としては、形態素解析の曖昧性、意味分類の曖昧性、格フレームの曖昧性などが考えられるが、以下では形態素解析の曖昧性削減について述べる。

学習データ：

走る　と　　彼　　は　　言う
(動詞) (引用助詞) (名詞) (助詞) (動詞)
走る　と　　彼　　は　　困る
(動詞) (接続助詞) (名詞) (助詞) (動詞)

「と」の品詞推定は？引用助詞／接続助詞？
‥ **走る　と　　彼　は**‥
(動詞)　(**？**)　(名詞)

図 2：最尤品詞推定での曖昧性削除が困難な例

3　形態素の曖昧性削減モデルの構築

3.1 最尤品詞推定による曖昧性削減の問題点

削除しても構文解析結果を変化させない形態素解析結果の曖昧性を判別するモデルを作成する方法としては、単純には、構文解析によって曖昧性が解消された品詞列を正解として、HMMやCRF[2]等により学習を行い、尤もらしい品詞を推定することでそれ以外の品詞の曖昧性を削除するという方法が考えられる。しかしながら、最尤推定を用いた方法では、統計モデルの性質上、誤って正しい形態素を削除してしまう可能性が残る。統計モデル学習においては、データスパースネスを防ぐために参照する情報をn-gramといった有限の空間に制限するため、その空間外の情報を用いないと解消できない曖昧性は本質的に正確な採否判定ができない。図２はbigramモデルの最尤推定において正確な曖昧性削除が困難な例である。このように学習データが対立する場合には、最尤推定ではより高頻度の解釈が選択されるため、低頻度の解釈が正しい場合には誤った曖昧性削除が起こる。

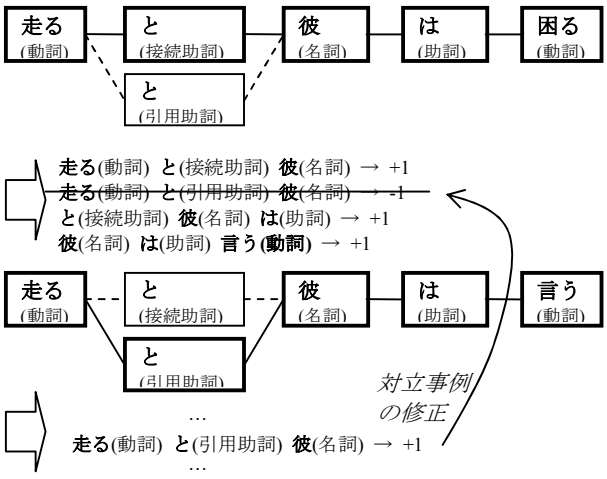

図 3：学習用事例の作成例
(太枠が構文解析結果で選択された曖昧性)

3.2 対立事例の修正（提案法）

以上を踏まえて、本稿では尤もらしい形態素列を１つ推定するのではなく、確実に不必要と思われる形態素のみを削除するというアプローチを取る。これにより、必要に応じて複数の曖昧性を残すことを可能とする。具体的には、図３に示すように、要不要を判別したい形態素を含む前後の形態素並びを素性とし、判別したい形態素の要不要をクラス(必要：+1, 不要：-1)とした分類学習を行い、得られたモデルを実行時に参照して不要な曖昧性を判別し、削除する。更に学習データに対立する事例（素性が同一でクラスが異なる）が存在する場合は、不要クラスに分類されている事例を学習データから削除する。この修正を施した学習データから得られるモデルは、曖昧性解消が本質的に困難な曖昧性を無理に解消することによる形態素の削除誤りを減少させる。

4　SVMの処理の効率化

今回、学習器としては、汎化能力が高く、効率よく素性の組み合わせを考慮しながら学習が可能なSVMを用いた。ただし、SVMを用いた学習では、学習コーパスが増えるごとに学習時間、判定時間が共に大きく増加してしまう。学習時間の増加は十分な量の学習コーパスを扱えないことによる精度の低下に繋がり、判定時間の増加は、処理の高速化という点から望ましくない。そこで、以下のような工夫により、学習コーパスの量を抑えている。

[対立事例の追加検出]

精度を高めるためには学習コーパスを増やす必要があるが、学習時間の問題から学習コーパスの増加には上限がある。そこで、学習コーパスとは別にコーパスを用意し、学習データ中の対立事例の検出を新たなコーパスから得られる事例も用いて行う。学習コーパスのみからは検出できない本質的に曖昧性解消が不可能な事例を新たに検出し、修正することができるため、学習コーパスの量は増やさずに削除誤りを減少させることができる。

[誤り駆動学習]

まず学習データのごく一部のみを用いて学習を行う。残りの学習データに関しては、現状の判別モデルで判定に誤った事例のみを追加してゆき随時モデルの更新を行う。これにより、同等の性能を持つモデルをより少ないSupport Vector数で実現できる。

表 1： 使用した素性

判別対象語、その前後の語	見出し、品詞、左接続番号、右接続番号
判別対象語	裏の全曖昧性の品詞

[品詞毎にモデルファイルを分割]

必要性の判別を行う語の品詞に関して判別モデルの独立性が高いことを利用して、判別する語の品詞ごとにモデルファイルを分割して学習・判別を行うようにした。これにより、精度をあまり劣化させることなく、学習時間、判別時間共に大幅に削減できる。

5 実験・考察

5.1 実験設定

提案手法の性能を調べるために、評価実験を行った。以下の実験では学習に日経新聞コーパス[4]93年の後半の68万文、評価用に日経新聞コーパス94年よりランダムに抽出した1万文を用いた。SVMの学習ツールにはTinySVM[5]を、Kernel関数は多項式カーネルを用いた。Kernel関数の次元数は比較実験の後、最も高い精度が得られた3を用いている。

表１に学習に用いた素性を示す。必要性の判別を行う語とその前後の語の単語情報と、判別を行う語の裏の曖昧性の品詞を素性として用いた。ただし、品詞体系は我々の言語解析系独自のものであり、左/右接続番号とは、各単語に付与された形態素接続検定用の情報である。活用型、活用形の情報はここに含まれる。左右接続番号との依存関係が強い一部の品詞では、左右接続番号を利用していない。また各形態素の見出しに関しては、現在の学習量では過学習に陥って誤った曖昧性削除が増加する傾向が見られたので使用していない。また、２つ前、２つ後の形態素に関しても同様の理由で使用していない。

5.2 実験結果

実験の結果を表２に示す。縦軸は実験環境を表し、全品詞が全ての品詞を判定対象としたとき、４品詞が削減効率のよい４品詞のみを判定対象としたとき、従来が本手法を適用しなかったとき、クローズドは評価用コーパスで学習したとき、の結果を示している。削減効率のよい品詞の選定方法は、図４のように品詞ごとに誤り数とエッジ削減数をプロットし、分離効率の良い（図の右下に分布している）４品詞を選択した。

表 3：曖昧性削除の具体例

表現	削除結果の説明
千葉銀行は借り入れ期間中に…	「借り入れ」の連用中止が消え、派生名詞のみに
福岡・白紙調書疑惑…	「・」の一般記号が消え、連結記号のみに
９５年の九州の主な予定。	２つめの「の」の独立助動詞が消え、格助詞のみに
職を求める学生が…	「求む」の可能動詞が消え、動詞「求める」のみに
中軸で安定した働きが見込める…	「働き」の派生名詞が消え、普通名詞のみに

解析時間は従来を100とした時の相対値である。結果、４品詞での曖昧性削減時に、従来と比較してエッジ数比で４％、解析時間比で2.4%ほどの削減となった。

表３に実際に削除された曖昧性の例を示す。

5.3 考察

5.3.1 曖昧性削減効率について

表２を参照すると、クローズドの評価では、エッジ数比で曖昧性を15%程カットできているので、潜在的にはより大きな割合での曖昧性削減も可能と思われる。

表 2：実験結果

	エッジ数	解析時間 (相対値)	誤削除 形態素数
全品詞	6326万	102.1	398
４品詞	6545万	97.6	94
従来	6818万	100	—
クローズド	5373万	91.6	1

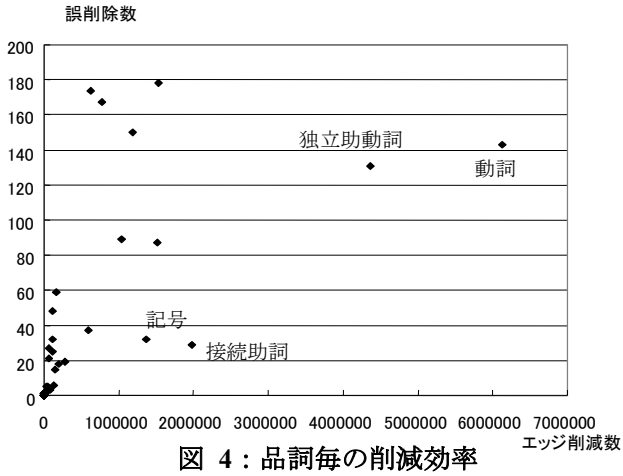

図 4：品詞毎の削減効率

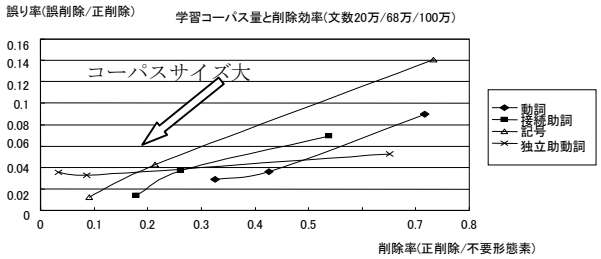

図 5：学習コーパス量と削減精度・削減効率

曖昧性削減率が低かった理由は、今回はデータスパースネスの問題から見出しを一切用いず、参照する文脈も前後１単語ずつと狭いために十分な分離性能が得られなかったものと思われる。今後は、何らかの素性選択アルゴリズムを応用することで、見出しを素性に含めてもデータスパースネスに陥らないような工夫を行って分離性能を高めていきたいと考えている。全品詞での曖昧性削減時には従来よりも解析時間が延びているが、これは判別処理のオーバーヘッドが高速化性能を上回ってしまったからである。今後更に学習データを増やすことを考えると、効率のよい素性選択の導入ないし高速な判定を行えるSVM以外の学習器の利用も検討したい。

また、形態素の削除誤り率について、表４にまとめた。削除対象数は、削除可能な形態素のうち判定対象の品詞の形態素数、正削除数は正しく削除した形態素数、誤削除数は削除誤りを表す。４品詞での曖昧性削減時の誤削除数が削除対象数25183に対して94と少ないことは、対立事例の修正の効果を表している。

最後に、削減効率のよい４品詞について、学習コーパス量と曖昧性の削減効率、削除精度の関係を図４にまとめた。特に接続助詞と記号について、学習コーパスを増やすに従って、一定の削除率を残しながら誤り率が0に近づいているのが分かる。

5.3.2 SVMの処理の効率化に関して

４章で述べたSVMの処理の効率化の各々についても実験を行い、効果を確認した。

[対立事例の追加検出]

新聞１万文での学習時に20万文の追加コーパスを対立事例追加検証用に用いた。結果、削除誤りを10%低減した。なお、この対立事例の追加検出は、劇的な効果こそないが、追加検証用のコーパスの量に上限はなく、また追加コーパスを解析する時間以外のデメリットはない。

[誤り駆動学習]

20万文の学習コーパスを1万文ずつに分割し、誤り駆動学習を行った。結果、通常の学習方法によるモデルと比較して、およそ半分のSupport Vector数でほぼ同等の性能のモデルを獲得することができた。

表 4：形態素の削除誤りの詳細

曖昧性も含めた全形態素数	331959		
うち、削除可能な形態素数	66837		
	削除対象数	正削除数	誤削除数
全品詞	66837	10896	398
４品詞	25188	3338	94

[品詞毎にモデルファイルを分割]

品詞毎にモデルを分割学習することで、新聞1万文での学習時間が、３時間から30分に短縮された。判別精度には変化がなかった。またモデルファイルを分割したことにより、判別効率の良い品詞だけを選り分けた新しいモデルを作成することもできた。

6 おわりに

本稿では、構文解析システムの辞書や文法、スコアリング関数を固定した場合に不必要となる曖昧性を判定するモデルを予め作成しておくことで構文解析処理を高速化する手法を提案した。本手法を形態素解析の曖昧性削減に適用した実験では、一定の曖昧性削減効果を示せたものの、十分な曖昧性削減が行えないなどの問題が残った。今後は、素性選択アルゴリズムを利用することで、より広い文脈をより効率的に素性として用いてモデルを詳細化することで精度向上を図っていきたいと考えている。またモデルを複雑化していった場合、SVMの判定時間の遅さがネックとなり高速化できなくなる可能性が高いため、他の判定アルゴリズムの使用を検討する。

参考文献

[1] 山端，安藤，三村：語彙化されたツリーオートマトンに基づく会話文翻訳システム 言語処理学会第6回年次大会(2000)

[2] 工藤，山本，松本：Conditional Random Fields を用いた日本語形態素解析，情報処理学会自然言語処理研究会 SIGNL-161(2004)

[3] 工藤，松本:カーネル法を用いた言語解析における高速化手法，情報処理学会論文誌 Vol.45 No.9(2004)

[4] 日経全文記事データベース http://sub.nikkeish.co.jp/gengo/zenbun.htm

[5] TinySVM http://chasen.org/~taku/software/TinySVM/

[6] Vapnik, V. : The Nature of statistical Learning Theory. Springer Verlag (1995)