5-4　基盤的言語処理ツール

5-4 Fundamental Natural Language Processing Tools

風間淳一　　王　軼謳　　川田拓也
KAZAMA Jun'ichi, WANG Yiou, and KAWADA Takuya

要旨

本章では、情報分析研究室で研究開発を行い、高度言語情報融合フォーラム（ALAGIN）から公開している基盤的言語処理ツール（評価情報分析器、形態素解析器、構文解析器）について解説を行う。

In this paper, we describe the fundamental natural language processing tools (evaluative expression analyzer, morphological analyzer, and syntactic parser) that we have developed and released through Advanced Language Information Forum (ALAGIN).

[キーワード]
自然言語処理，評価情報分析，形態素解析，構文解析，高度言語情報融合フォーラム（ALAGIN）
Natural language processing, Evaluative expression, Morphological analysis, Syntactic analysis, Advanced Language Information Forum (ALAGIN)

1　まえがき

自然言語で書かれた文書から有用な情報や知識を抽出し、様々なアプリケーションで利用するためには、まず、文を計算機が（意味を理解して）処理し易い形式に変換する必要がある。これらの変換処理のうち、有用性に一定のコンセンサスが得られている処理をここでは基盤的言語処理と呼ぶ。代表的なものには、文を単語に区切って品詞を付与する「形態素解析」、語の間の係り受け構造などを決定する「構文解析」などがある。また、最近では、本章でも紹介する、ある表現が肯定的な意見を表明しているのか、否定的な意見を表明しているのかを判定する評価情報分析も基盤的な処理として定着してきている。我々は、これらの基盤的言語処理についての研究開発を行っている。加えて、開発したシステムをオープンソースとして ALAGIN を通して一般に公開することで、成果の社会還元を積極的に行っている。まず、**2** では、我々の評価情報分析システムについて解説を行う。これは情報信頼性プロジェクトにおいて開発された情報分析システム WISDOM（http://wisdom-nict.jp/）でも使用されている技術で、我々は、この技術を整理し、辞書などを整備してオープンソースソフトウェアとして一般公開を行っている。形態素解析、構文解析は比較的古い研究分野であり、日本語に関しては十分な精度を持つ解析器が一般公開されて既に広く利用されているが、中国語など他の言語に関しては研究の歴史がまだ浅く、近年盛んに研究が行われるようになっているが精度は不十分である。今後、重要性を増す多言語の処理に対応するため、我々は、特に中国語に焦点をおいて研究開発を行い、世界的にトップレベルの精度をもつシステムを開発してきた。**3**、**4** では、これらの中国語用形態素解析器と構文解析器について解説を行う。

2　評価情報分析システム

テキストから人々の意見や評価を抽出する評価情報分析技術が、近年注目を集めている。評価情報分析では、与えられた文が何らかの対象に対する意見や評価としてみなせるかどうか判定し、それが肯定的な意見なのか否定的な意見なのか、といったことを自動的に判定する。評価情報処理技術が注目されている背景として、Web を始めとする情報媒体の発達が挙げられる。Web によって多くの人々が、様々な話題について自分たちの

意見や評価を公に発信できるようになった。そのため、人々の意見や評価は日々大量に蓄積される一方で、大量の意見、評価を集約するために効率的にそれらを抽出し、分類する技術の研究が求められてきている。そこで本稿では、我々が研究開発を行っているテキストから肯定的もしくは否定的な意見や評価を自動的に抽出し、分類する評価情報分析システムについて報告する。

2.1 評価情報

意見や評価は様々な形で表明される。本稿では、評価情報とはテキスト中から何らかの対象に対する肯定（否定）的な判断や態度が読み取れる情報とする。より具体的には、評価情報は、「評価をする者（評価保持者）」、「評価の対象（評価対象）」、「言語によって表現された判断や態度（評価表現）」、「評価タイプ」と「評価極性」の五つの要素を基本構造として持つ情報とする。例1では「太郎」による「青森のリンゴ」についての肯定的な感情が記述された文と解釈することができる。このとき、「大好きだ」は実際に評価として読み取れる言語表現なので「評価表現」として抽出される。また、その評価を下している「太郎」は評価保持者として抽出され、「青森のりんご」について評価しているので、それが評価対象として抽出される。以降では、具体的な評価情報を提示する際には、評価対象には下線を付与し、評価表現は太字で表すこととする。評価保持者は多くの場合、文の著者自身となるため、その場合は、文に明示的に評価保持者が書かれない事が多い。評価保持者が明示的に文に現れる場合は斜体で表すこととする。

例1：*太郎*は　青森のりんごが　**大好きだ。**
　　評価保持者　評価対象　評価表現（感情＋）

実際のテキストにおいて、評価表現は、感情に基づくものや、経験等事実に基づくものなど様々な表現で述べられる。そこで我々は評価表現の意味や極性の有無などの観点から次のように分類した（＋は肯定的な評価極性、－は否定的な評価極性を表す）。

(1) 感情＋、感情－：主観的でかつ、感情的な評価表現
　例2：京都が**好きだ。**（感情＋）
　例3：*太郎*はA製品には**興味ありません。**（感情－）
(2) 批評＋、批評－：主観的ではあるが、賛成／反対・称賛／批判等の態度を表す評価表現
　例4：京都は**美しい。**（批評＋）
　例5：A制度には**問題がありすぎる。**（批評－）
(3) メリット＋、メリット－：長所や欠点について記述された評価表現
　例6：このクーポンは**いつでも利用できます。**（メリット＋）。
　例7：A製品は**使いにくい。**（メリット－）
(4) 採否＋、採否－：積極的に行為や利用を却下したり、促したりする行為を表す評価表現
　例8：*A社*は電子マネーの**投入を決定した。**（採否＋）
　例9：A製品は**人気がありません。**（採否－）
(5) 出来事＋、出来事－：良い／悪い出来事や経験を表す評価表現
　例10：A製品は**グッドデザイン賞を受賞した。**（出来事＋）
　例11：B製品は**買って三日後に壊れてしまいました。**（出来事－）
(6) 当為：義務や提言、対策を表す評価表現
　例12：電子マネーを**投入するべきだ。**（当為）
　例13：裁判員制度は**国民の理解を得た上で進めていくべきだ。**（当為）
(7) 要望：要望や希望を表す評価表現
　例14：電子マネーを**使えるようにしてほしい。**（要望）

当為や要望においては、例13のように、特定の対象（この場合は「裁判員制度」）について肯定（否定）的な判断が必ずしも明確に示されない場合があるため、極性は付与しないこととした。

2.2 評価情報コーパス

従来抽出することが困難であった多様な評価情報の抽出を実現するために、評価情報の付与されたコーパスを作成した[1]。このコーパスは、「電気自動車」や「年金問題」などの100個のトピックに対して、各トピックについて200文ずつ、合計20,000文をWeb上の文書から収集して作成したコーパスである。このコーパスでは、

2.1 で述べた評価情報が付与されている。抽出された評価情報が与えられたトピックに関連している評価情報かどうかを表す情報が付与されている。例えば、「裁判員制度」というトピックに対して、「このサイトでは裁判員制度について興味深い考察が書かれている」というような文が与えられた時、この文は、「裁判員制度」自体に対して評価するわけではない。むしろサイトについての評価として読み取れる。このようにトピックに対する評価にはつながらない評価情報に対しては、その評価情報がトピックとは関連のないものであることを示す情報が付与されている。このコーパスは、機械学習の訓練データとして使用したり、性能評価のためのテスト用データとして使用される。

2.3 評価表現辞書

評価表現辞書とは、評価表現とその表現が持つ評価極性の組（例:「規則正しい　＋」「甘ったるい　－」など）の集合である。この辞書は評価情報分析において基礎的な知識として利用される。ここでは、次に述べる手法を用いて辞書を構築した。まず、評価極性が既知である少数の評価表現を種となる表現として用意する。そして、文脈類似語データベース[2]、カスタム単語集合作成ツール[3]（いずれも、意味的に類似する単語の集合を作成することができる）を利用して、種の表現と意味的に類似する語は評価表現である可能性が高いという仮定のもと評価表現候補を作成する。その評価表現候補に対して、評価極性の有無を手作業で判定し、評価極性を持つ評価表現をその評価極性と共に辞書に登録する。上記の過程を繰り返し、ブートストラップ的に種表現から評価表現を順次増やしていく。さらに、負担・トラブル表現リスト[4]の見出し語も「－」の極性を持つ評価表現として登録した。辞書中に登録されている評価表現の数は合計で 36,981 個である。なお、この評価表現辞書は ALAGIN において「意見（評価表現）抽出ツール用モデル」の一部として公開されている。

2.4 評価情報の抽出

2.4.1 評価情報抽出の流れ

評価情報分析システムにおける評価情報抽出の流れを図 1 に示す。はじめに利用者からテキストが入力されると、入力テキストから評価表現の抽出が行われる。続いて、評価保持者の同定、評価タイプの分類と評価極性の分類が行われる。最後に結果が出力される。以下の節では、各処理について説明する。

2.4.2 評価表現の抽出

評価表現の抽出手法としては、条件付き確率場（Conditional Random Field、CRF）により文中の各形態素に評価表現の開始（B)、中間（I)、評価表現以外（O）を表すタグを付与する方法[5]を用いている。これは固有表現抽出等の情報抽出で良く用いられる方法である。ここで抽出対象としている評価表現は文中の任意の箇所に出現する可能性があるため、このような系列ラベリングの手法を用いることにした。評価表現の抽出を行う際に、評価を表すためによく使用される単語の情報は非常に有用であると考えられる。そこで、前述した評価表現辞書を用い、CRF の素性としては、前後 2 つまでの形態素の出現形、原形、品詞大分類、品詞細分類、評価極性語辞書中での極性を使用する。

2.4.3 評価保持者の同定

評価保持者の同定は、2 つのステップにより行っている。はじめに、与えられた評価表現に対して、その評価保持者がその著者と同一であるかどうかを SVM（Support Vector Machine）を用いて判定する。素性としては、評価表現に含まれる形態素の出現形、原形、品詞大分類、品詞細分類を用いる。もし著者と同一ではないと判定された場合は、CRF を用いてその評価表現が含まれる文中から評価保持者を抽出する。その際の素性としては、各形態素の出現形、原形、品詞大分

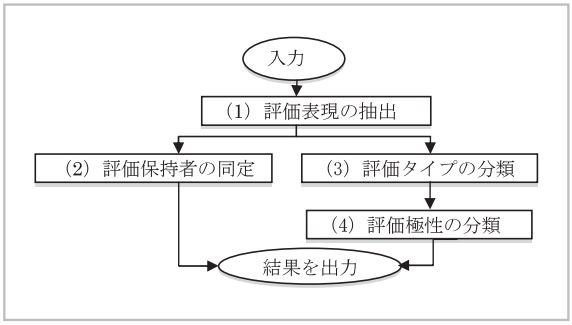


図 1 評価情報抽出の流れ

類、品詞細分類、評価表現との位置関係を用いる。

2.4.4 評価タイプの分類

評価タイプの分類では、与えられた評価表現が**2.1**の7種類の評価タイプのいずれであるかペアワイズ法を用いて多値分類に拡張したSVMを用いて判定する。素性としては、評価表現に含まれる各形態素の出現形、原形、品詞大分類、品詞細分類とそれらの組み合わせを用いる。

2.4.5 評価極性の分類

評価極性の自動分類については、これまでに様々な研究が行われている[6][7]。評価極性分類の代表的なアプローチとして、Bag-of-Words素性を用いた教師あり機械学習を適用する方法がある。この方法は、評価表現をそこに含まれる単語の集合として表現し、その評価極性を分類する手法である。しかし、評価極性の分類では、評価極性の反転がしばしば起こる。「ガン細胞を消滅させる」という評価表現の場合、「ガン細胞」自体は否定的な意味を持つ単語であるが、「消滅」という単語に係ることで極性が反転し、全体としては肯定的な意味を持つ。このように評価極性の分類では肯定的（または否定的）な単語が出現していても、それが評価表現全体の極性と等しいとは限らないため、評価表現中の個々の単語を独立に扱うのではなく単語間の相互作用を考慮する必要がある。そこで我々のシステムでは、そのような相互作用を考慮することができる「隠れ変数を持つ条件付き確率場」を用いた評価極性分類手法を利用している[8]。この手法では評価表現の依存構造木を考え、個々の部分依存構造木に対する評価極性を隠れ変数で表し、隠れ変数間の相互作用を考慮して評価極性分類を行う。

例として「不安やストレスを減らす効果がある」という評価表現を考える。この文では、「不安や」や「ストレスを」という文節自体は否定の極性を持つが、それらの文節が「減らす」という文節に係ることで評価極性が反転し、「不安やストレスを減らす」という部分依存構造木は肯定の極性を持つと考えることができる。また、「不安やストレスを減らす効果が」や「不安やストレスを減らす効果がある」という部分依存構造木の極性も肯定である。このように、評価表現の依存構造木の各部分木に対して評価極性を考えることができる。

そこで、図2のグラフで示されるような確率モデルを考えることにする。この確率モデルでは、評価表現の各文節が確率変数を持つものとする（図2では丸いノードで表されている）。この確率変数は、その文節をルートとする部分依存構造木の評価極性を表す。この確率変数は、その文節に含まれる単語の影響を受けるだけではなく、依存関係にある文節の確率変数に対しても相互に影響を受けるものとする。このようなモデルを利用することにより、肯定的（または否定的）な文節は肯定（または否定）の極性を持ちやすいという情報や、係り先の文節に極性を反転させる単語が含まれる場合は係り元と係り先の文節の極性が逆になりやすいといった情報を表現することができる。実験の結果、評価表現を単純な素性の集合として表現して分類する手法と比べ、本手法は高い分類精度を達成することが確認されている[8]。

2.5 性能評価

2.2で説明した評価情報コーパスを用いて、評価情報分析システムの性能評価を行った。コーパスはランダムに等分割し、10分割交差検定を行った。各モジュールは単体で独立して動かして評価を行った。評価表現の抽出については、正しく抽出された評価表現の数を正解データ数の評価表現の数で割った値である再現率、正しく抽出された評価表現の数をシステムが出力した評価表現の数で割った値である適合率、再現率と適合率の調和平均であるF値により評価した。その際に、正解データ中の評価表現とシステムが出力した評価表現は、その主辞（主要な意味を表す語。日本語の場合は、末尾の形態素）が一致していれば評価表現が一致するとみなして評価を行っている。評価保持者の同定、評価タイプの分類および評価極性の分類については、テスト事例の中で正しい

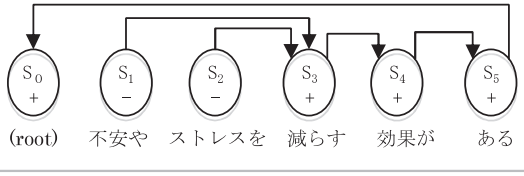


図2 部分依存構造木の評価極性の例

出力が得られた事例数の割合である正解率により評価した。評価情報分析システムの性能を表1に示す。

表2に、評価表現抽出の難易度の参考となる人間の作業者による一致に関する統計を載せる。評価情報コーパスを人手で作成した際には、コーパスの品質を高めるため、同じ文に対して2名の作業者がアノテーションをしている。表2は、その時の一方の作業者の結果を正解とみなし、もう一方をシステムの出力とみなした場合の再現率、適合率、F値である。この数値をみると、評価表現の抽出は高い一致率を得ることが困難なタスクであり、表1で示したシステムの性能はそれほど悪いものではないと言える。また、評価極性分類については、**2.4.5**で説明した隠れ変数をもつ条件付き確率場の方式と**2.3**で紹介した辞書を用いることで、0.87という高い正解率を実現している。

2.6 ALAGINでの公開

本システムは、ALAGIN言語資源サイトを通じてオープンソースソフトウェアとして配布している（http://alaginrc.nict.go.jp/opinion/index.html）。また、評価情報分析システムのモデルパラメータ（プログラムの動作を規定する単語群や数値群）を格納したデータベースをALAGINから提供している。これには、システムの処理の流れに応じて「評価表現抽出」「評価保持者同定」「評価タイプ分類」「評価極性判断」の4種類のモデルファイルが含まれている。

表1 評価情報分析システムの性能

評価表現抽出	再現率	0.4077
評価表現抽出	適合率	0.6020
評価表現抽出	F値	0.4860
評価保持者同定	正解率	0.6919
評価タイプ分類	正解率	0.6515
評価極性分類	正解率	0.8703

表2 評価表現抽出に関する人間の作業者の一致率

再現率	0.67
適合率	0.71
F値	0.69

3 高精度中国語形態素解析システム

本研究では、大規模なラベルなしデータを利用し、中国語の形態素解析精度を向上させる、いわゆる半教師あり学習に基づく手法を提案した。より具体的には、ベースラインモデルを用いて大規模ラベルなしデータを自動解析して得られるN-gram情報、単語クラスタリングによって得られるクラスタ情報、交差検定法によって得られる辞書マッチング情報を追加的な素性として利用する。標準的な評価データであるPenn Chinese Treebankを用いた実験では、提案手法が半教師あり学習を用いないベースラインおよび既存手法より高い解析精度を達成することを示した。

中国語には日本語と同様に単語と単語の間に空白を入れる「分かち書き」という習慣がないため、形態素解析（単語分割と品詞タグ付け）は、中国語処理において最も基本的かつ重要な課題であり、構文解析や情報検索を始めとした多くのアプリケーションにおいて前処理として使用されるため、高い精度が必要である。中国語形態素解析に関しては近年様々な研究が行われている。特に最近は、単語分割と品詞タグ付けの同時学習が多く報告されている[9]-[13]。例えば、我々は単語一文字はハイブリッドモデルを処理方式として採用し、最高水準の解析精度を達成した[11]。

また、システムの性能をさらに改善するために、正解が付与されていない大量のデータを利用する、いわゆる「半教師あり学習」も盛んに用いられるようになってきている。既存研究の報告によれば、半教師あり手法を用いることで、いくつかの自然言語処理タスクで性能が向上することが示されている。例えば、テキストチャンキング[14]、品詞タグ付けと固有表現抽出[15]、係り受け解析[16]-[18]などでその効果が示されている。しかしながら、半教師あり手法を中国語形態素解析に利用した研究はこれまであまり行われていない。持橋ら[19]は半教師あり手法で中国語の単語分割精度を向上させたが、使用したラベルなしデータの規模が小さく、その差は僅かであった。

本研究では、同時学習よりも実装が容易なパイ

プラインシステムにおいて、大規模なラベルなしデータを利用することで、単語分割と品詞タグ付けの精度を向上させる方法を提案する。

3.1 システムの概要

我々のシステムは、開発コストを抑えることを1つの目標とし、実装しやすい2段階のパイプラインシステムを採用している。単語分割には文字ベースのCRFを用い、品詞タグ付けには単語ベースのCRFを用いる。CRFの実装としてはオープンソースのCRF++(version 0.54)[*1]を使用する。ベースラインの単語分割モデルでは、素性として、前後1つまでの文字、記号かどうか、文字タイプを使用する。「S（1つ文字の単語)、B（単語の最初)、B_2（単語の2つ目の文字)、B_3（単語の3つ目の文字)、M（単語の他の中間文字)、E（単語の最後)」を表す6つのタグを付与する。ベースラインの品詞タグ付けモデルでは、前後2つまでの単語、最初の文字、最後の文字、単語の長さを素性として使用する。

形態素解析システムを高精度化するために、ラベルなしデータの情報を新しい素性として導入するアプローチを提案する。最初に、ベースラインモデルを用いて大規模ラベルなしデータを自動解析する。次に、自動解析データから多様な辞書情報を抽出する。そして、これらの辞書情報を単語分割と品詞タグ付けの新しい素性として利用する。さらに、単語分割されたデータを用い、単語クラスタリングを行い、そのクラスタ情報を品詞付けの素性として導入する。さらに、交差検定法により、ラベルありデータから抽出された辞書情報も素性に加える。本手法の概要を図3に示す。以下の節では、新しい素性について説明する。

3.2 単語分割のための新素性

3.2.1 半教師あり N-gram 素性

ベースラインの単語分割モデルでラベルなしデータを単語分割し、分割された文から文字N-gramリストを抽出してN-gram素性を生成する。

ベースラインの単語分割モデルによって、ラベルなしの文の各文字 c_i にタグ t_i が与えられる。つまり、文字数 L とすると、自動分割の結果は系列 $\{(c_i, t_i)\}_{i=1}^{L}$ となる。この自動分割の結果からN-gramリスト $\{(g, seg, f(g, seg))\}$ が抽出される。ここで、g は文字N-gram（例えば、uni-gram c_i、bi-gram $c_i c_{i+1}$、tri-gram $c_{i-1} c_i c_{i+1}$ など）を表し、seg はN-gram g の分割プロフィールである。分割プロフィールはタグ t_i あるいはタグの組み合わせである（例えば、bi-gram $c_i c_{i+1}$ の場合は t_i あるいは $t_i t_{i+1}$ の形式で定義できる)。$f(g, seg)$ はN-gram g の分割プロフィールが seg である時の頻度である。

そして、その頻度によって、リストを高頻度(HF: トップ5%)、中頻度（MF: 5%から20%まで）と低頻度（LF: 残りの80%）の3つのセットに分ける。最後に、リスト $L_{ng} = \{(g, seg, FL(g, seg))\}$ が得られる。ここで、$FL(g, seg)$ は上述の方法で決めた頻度ラベルである。

N-gramリスト情報を新しい素性にエンコードするために、様々な素性表現を試したところ、$seg = t_i$ のbi-gramリストから得られる素性が最

*1 http://crfpp.sourceforge.net/

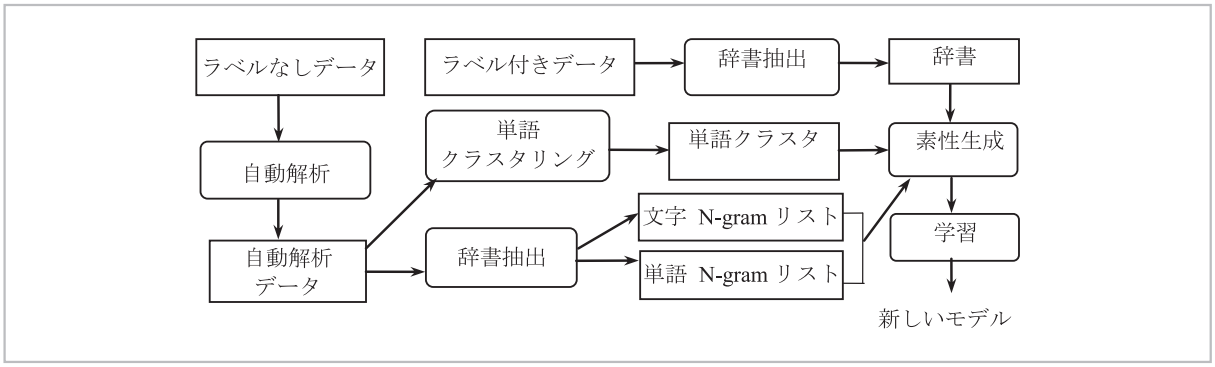


図3 提案手法の概要

も効果的であった。このリストを用い、現在の文字 c_0 に対して、次のように素性を生成する。L_{ng} から g が bi-gram c_0c_1 と照合できるサブセットを獲得し、このサブセットを L_m とする。L_m 中の各エントリーに対して、下記のような素性を生成する。

(a) ***seg-FL*(*g, seg*)**

そして、L_m 中の各エントリーの素性を1つの N-gram 素性として連結する。

例えば、L_m が {(*幸/福, B, HF*), (*幸/福, B*$_2$*, MF*), (*幸/福, E, LF*)} である。c_0c_1 =「*幸/福*」に対して、c_0 の N-gram 素性は「*B-HF*|*B*$_2$*-MF*|*E-LF*」である。

3.2.2 辞書素性

文字ベースの単語分割モデルは未知語の解析精度に優れている一方、既知語の解析精度が低いことが知られている。一般的に、既知語の解析精度は、辞書を用いることにより向上させることができる。既知語の辞書は、ラベルあり学習データから簡単に抽出することができる。そこで、本研究は辞書を利用した素性を導入することにした。この素性を「辞書素性」という。

学習データから単語と単語に対応するすべての品詞タグを集め、辞書を作成する。例えば、「交流」に対して、エントリーの内容は（交流 , *NN-VV*）である。ここで、「*NN-VV*」は学習データの中での「交流」に対応するすべての品詞タグを連結したものである。

しかし、学習データから抽出した辞書を用いて素性を生成して学習を行うと、辞書素性を過度に信用してしまうという学習データへの過学習が起きる。そこで、交差検定法の考え方を取り入れた下記の方法を用いて、辞書を構築し、使用する。

- 学習データを 10 個の等しいセットに分割する。
- 各セットに対して、残りの 9 セットを用い、辞書を構築し、この辞書を使用し、辞書素性を生成する。
- テストセットに対しては、学習データの全体を用い、辞書を抽出し、この辞書を用いて、辞書素性を生成する。

素性の生成の際には、辞書との前向き最左最長マッチを行い、単語を選ぶ。各単語 w の各文字 c_k に対して、下記の素性を追加する：

(b) ***P*(c_k)/*LEN*(w)-*POSs*(w)**

LEN(w) は単語 w の長さ、*P*(c_k) は文字 c_k が w 中の何文字目かを示す数、*POSs*(w) は単語 w の辞書中の品詞タグの組み合わせを表す。例えば、文字列 c_0c_1 =「幸/福」が辞書のエントリー「(幸福 , *JJ-NN-VA*)」と照合できた場合、c_0「幸」の辞書素性は「*1/2-JJ-NN-VA*」で、c_1「福」の辞書素性は「*2/2-JJ-NN-VA*」となる。

3.3 品詞タグ付けのための新素性

3.3.1 半教師あり N-gram 素性

ラベルなしデータを自動分割した結果を入力として品詞タグ付けモデルで解析すると、単語レベルの N-gram リスト L_{wg} = {(*w, pos, FL*(*w, pos*))} が得られる。ここで、w は単語 N-gram で、*pos* は単語 N-gram の品詞プロフィールである。この N-gram リストを利用し品詞タグ付けの N-gram 素性を生成する。予備実験によって、w が uni-gram で、*pos* が w の品詞である場合に、一番良い結果が得られることがわかった。L_{wg} から w が現在の単語 w_0 と照合できる照合エントリーを獲得し、このサブセットを L_s とする。例えば、w_0 が「研究」である場合に、照合エントリーは (研究 , *NN, HF*)、(研究 , *VV, HF*)、(研究 , *VA, LF*) と (研究 , *CD, LF*) などとなる。誤り分析によって、自動タグ付けによる誤りは問題になることが多いことが明らかとなったため、サブセット L_s を獲得する際、次のような制限を設けた。ここで、*N*(X) は *FL*(w, *pos*) = X となるようなエントリーの数とする。

i. *N*(*HF*) ≧ 2 の場合は、*FL*(w, *pos*) = *HF* である照合エントリーを L_s とする。
ii. *N*(*HF*) < 2 かつ *N*(*HF*) + *N*(*MF*) ≧ 2 の場合は、*FL*(w, *pos*) = *HF* と *FL*(w, *pos*) = *MF* である照合エントリーを L_s とする。
iii. *N*(*HF*) + *N*(*MF*) < 2 の場合は、すべての照合エントリーを取る。

例えば、上記の例「研究」において、L_s は {(研究 , *NN, HF*)、(研究 , *VV, HF*)} である。単語分割と同様に、L_s 中の各エントリーに対して、下記のような素性を生成する。

(c) ***pos-FL*(*w, pos*)**

そして、L_s 中の各エントリーの素性を1つの N-gram 素性に連結する。例えば、w_0 =「研究」に対して、w_0 の N-gram 素性は「*NN-HF*|*VV-HF*」

である。

3.3.2 半教師ありクラスタ素性

自動解析のデータを用い、単語クラスタリングを行う。Koo ら [18] の方法を参考にし、Brown クラスタリング法 [20] で得られるクラスタ階層の prefix を用い、様々な粒度のクラスタ素性を作る。予備実験の結果から、下記のクラスタ素性を使用することにした。

(d) $\mathbf{w}_{-1}$、$\mathbf{w}_0$、$\mathbf{w}_1$ の階層ビット表現の全ビット
$\mathbf{w}_{-1}$、$\mathbf{w}_0$、$\mathbf{w}_1$ の階層ビット表現の前 6 ビット

予備実験では、これらのクラスタ素性を bi-gram テンプレートとして使用した場合に最も精度が良かった。

3.3.3 辞書素性

単語分割と同じ辞書を使用し、素性を追加する。現在の単語 w_0 に対して、下記の素性を与える。

(e) $\boldsymbol{POSs}(\boldsymbol{w_0})$

$POSs(w_0)$ は辞書にある単語 w_0 の品詞タグを連結したものである。

3.4 実験

3.4.1 データセット

(1) ラベルありデータ

Penn Chinese Treebank を用い、実験を行った。具体的には、CTB5(LDC2005T01)、CTB6(LDC2007T36) と CTB7(LDC2010T07) を使用した。これらのコーパスは、表 3 に示すように、学習セット、開発セットとテストセットに分割して用いる。既存研究では CTB5 がよく用いられるが、CTB6 と CTB7 はテストセットと開発セットの規模が大きいため、パフォーマンスに及ぼす影響をより信頼性高く判断できる。

(2) ラベルなしデータ

Chinese Gigaword Version 2.0 (LDC2009T14) の XIN_CMN 部分から CTB と重複する恐れのあるデータを取り除いて、残りの 2.04 億語をラベルなしデータとして使用した。単語クラスタリングにはそのうち 100 万語を使用した。

表 3 実験用コーパス情報

	学習セットの文数	開発セットの文数	テストセットの文数
CTB5	18,089	350	348
CTB6	23,420	2,079	2,796
CTB7	31,131	10,136	10,180

3.4.2 実験結果

提案手法の有効性を評価するために、中国語の単語分割 (Seg) と品詞タグ付け (Seg & Tag) の実験を行った。精度の評価には、F 値を使用した。表 4 に CTB5 のデータを用いた先行研究の結果と本提案手法による結果を載せる。先行研究の結果は全て論文から引用したものである。本提案手法は単語分割も品詞タグ付けも最も良い精度を達成している。

さらに、CTB6 と CTB7 を用い、Kruengkrai ら [10] と Kruengkrai ら [11] に述べられている方法との比較実験を行った。本提案手法による結果との比較を表 5 に示す。より大きいデータセットを用いて評価した場合でも本提案手法が最高精度を達成していることが分かる。

3.5 システムの公開

本システムは、「CSP (Chinese Word Segmenter and POS Tagger)」という名称で、ALAGIN の言語資源サイト (http://alaginrc.nict.go.jp/csp/index.html) を通じてオープンソースソフトウェ

表 4 先行研究との比較 (CTB5)

Method	Seg	Seg & Tag
提案手法	**0.9812**	**0.9420**
ベースライン	0.9753	0.9318
Zhang ら [9]	0.9778	0.9367
Kruengkrai ら [10]	0.9787	0.9367
Kruengkrai ら [11]	0.9798	0.9400
Jiang ら [12]	0.9785	0.9341
Nakagawa ら [13]	0.9796	0.9338

表 5 先行研究との比較 (CTB6 と CTB7)

	CTB6		CTB7	
Methods	Seg	Seg & Tag	Seg	Seg & Tag
提案手法	**0.9579**	**0.9113**	**0.9566**	**0.9051**
ベースライン	0.9513	0.8999	0.9498	0.8937
Kruengkrai ら [10]	0.9550	0.9050	0.9540	0.8986
Kruengkrai ら [11]	0.9551	0.9053	0.9546	0.8990

アとして公開予定である。同時に、モデルパラメータ（プログラムの動作を規定する単語群や数値群）を格納したデータベースを ALAGIN から提供する。データベースには CTB5、CTB6 と CTB7 で学習されたモデルと、対応する N-gram リスト、クラスタリングの情報などが含まれている。

4　高精度中国語係り受け解析

形態素解析の後には、通常、文の構造を決定する構文解析と呼ばれる処理が行われる。構文解析の中でも近年盛んに研究されているのが、動詞とその主語や目的語単語間の関係（係り受け）を決定する係り受け解析と呼ばれている処理である。ここでは、我々が開発した半教師有り学習を取り入れた高精度な係り受け解析器 [21][22] について解説する。このシステムは中国語に関して世界最高性能を達成している。

図 4 は、「布朗一行于今晚离沪赴广州。（ブラウン一行は今夜上海を離れ広州に向かう。）」という中国語の文を形態素解析し、さらに、係り受け解析する様子を表している。係り受け関係は矢印（弧）で表され、矢印の元の語が先の語へ「係る」と表現する。弧にはその種類（主語を表す subj、目的語を表す obj など）を表すラベルが付与されることもある。ROOT は文の主要な動詞の位置を表すための架空の単語である。全体として、ROOT を根とする木構造となる。中国語の場合には、係り受け解析結果の弧は単語を図のように文に出現する順に一列に並べたときに交差しないという制約がある。日本語の場合には、さらに、必ず前から後ろへ係るという制約がある。日本語や中国語などでも、特定の例外では交差したりする可能性もあるが、交差しないと仮定して処理を効率的に行うことがしばしば行われる*2。

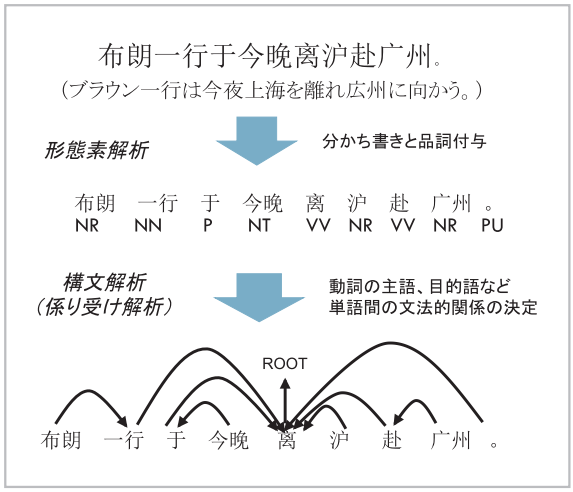


図 4　中国語の係り受け解析の様子

係り受け解析を行う様々な手法が提案されているが、その精度の良さから、グラフベースの手法 [23][24] が近年広く用いられるようになっている。この手法では、文中の各単語をノードとみなし、各ノード間を両方向の弧が結んでいるようなグラフを考えて、このグラフの全域木（すべてのノードを含み、木となっている部分グラフ）の内、（非交差制約がある場合にはそれを満たし）最大の重みをもつもの（Maximum Spanning Tree）を見つけることで係り受け解析を行う（MST パージング）。重みは、各弧に重みが設定される場合（1 次モデル）[23]、加えて 2 つの弧に対して重みが設定される場合（2 次モデル）[24] など、様々なバリエーションがあり、これらの各重みを全域木全体で和をとったものが、全域木の重みとなる。なお、次数（重み設定に同時に関わる弧の数）が上がるほど、処理のコストは大きくなるため、通常、上で述べたような 1 次や 2 次のモデルがよく用いられる。我々も、ここでは 1 次モデル [23] と 2 次モデル [24] を使用した。各重みは、さらに、単語やその組み合わせなど様々な素性関数に対する重みとして分解される。例えば、1 次モデルの場合には、以下のようになる。

$$w(x,y)=\sum_{(i,j)\in y} w(i,j)=\sum_{(i,j)\in y} \alpha\cdot f(x,i,j)$$

ここで、x は入力単語列、y は全域木である。(i,j) は、i 番目の単語から j 番目の単語への弧を表す。$f(x,i,j)$ i は、弧 (i,j) の様々な特徴を表した素性ベクトルであり、α は各素性の重みを表す重みベクトルである。重みベクトル α は、機械学習手法により人手で作成した正解データから自動で獲得される。

*2　なお、チェコ語など交差が頻繁におこる言語もあり、その場合には交差を許すモデルが用いられる。

4.1 部分木素性の利用

このシステムでは、解析精度を改善させるため、半教師有り学習を取り入れた手法を用いている。半教師有り学習とは、通常の人手による正解データに加えて、大量の生文（生コーパス）のデータを用いて精度を向上させるような手法のことを言う。ここでは、正解データから学習した1次のMST解析器（ベースラインモデル）を用いて大量の生文を係り受け解析し、その結果から、1次と2次の部分木を抽出する。さらに、それらの部分木を出現頻度により、HF（高頻度：頻度上位10%）、MF（中頻度：次の10%）、LF（低頻度：それ以外）とZERO（ゼロ：1回も出現しない）に分類して、この分類ラベルを、係り受けの際の素性の1つとして利用する（詳細は、論文[21]を参照）。直感的には、ベースラインモデルの結果には誤りも含まれるが、解析が難しい文ばかりではないため、統計をとれば良く係りやすい語の組や、逆にほとんど係らない語の組の傾向が分かり、その情報が、正解データを用いた学習の際にうまく利用できるということである。

図5は、解析結果からの部分木の抽出の様子を表している。なお、論文[21]で用いた2次モデル[23]では、隣接する2つの弧のみを許すため、抽出される2次の部分木もそのように制限されている。論文[22]では、より高度な2次モデル[25]を利用して、「親–子–孫」という形の2次の部分木など利用できるようにしている。

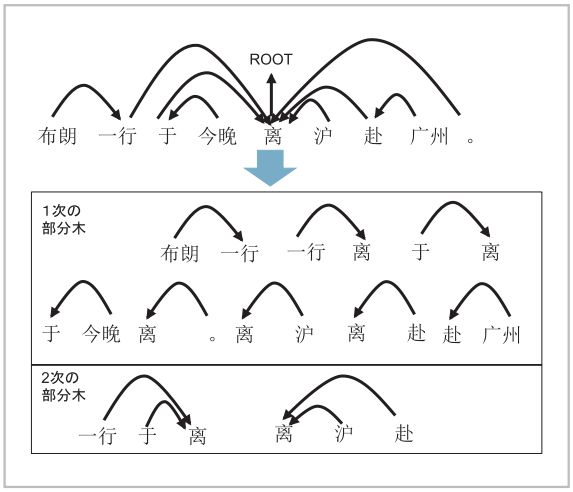


図5 部分木の抽出

4.2 実験

英語と中国語で提案手法の評価を行った。ここでは、文献[22]に基づいて結果を示す。英語では、標準的な学習・評価データであるPenn Treebankを使用し、生コーパスとしては、4,300万語からなるBLLIPコーパスを用いた。中国語でも、やはり標準的な学習・評価データであるChinese Penn Treebank（Version 4.0）を用い、生コーパスとしては、3.11億万語からなるChinese Gigawordコーパス（Version 2.0）を用いた。評価は、各語（句点を除く）の係り先を正しく決定できた割合（UAS：Unlabeled Attachment Score）と、一文のすべての係り受けが正解と完全一致した文の割合（Complete）という指標で行った。表6に英語、表7に中国語の結果を示す。両言語とも、部分木素性を用いることで、大幅に精度が向上することが分かる。また、クラスタリング素性[26]や、他の解析器の出力を利用する統合素性[27]などと同時に使用することで、更なる精度向上が可能である。既存研究との

表6 実験結果（英語）

	UAS	Complete
1次モデル	90.95	37.45
1次モデル＋部分木素性	91.76	40.68
2次モデル	91.92	44.28
2次モデル＋部分木素性	92.89	47.97
2次モデル＋部分木素性＋クラスタ素性＋統合素性	93.55	49.95
KOO08-dep2c [6]	93.16	N/A
Carreras 2008 [8]	93.5	N/A
Suzuki 2009 [29]	93.79	N/A

表7 実験結果（中国語）

	UAS	Complete
1次モデル	86.38	40.80
1次モデル＋部分木素性	88.11	43.10
2次モデル	88.59	48.85
2次モデル＋部分木素性	91.77	54.31
2次モデル＋部分木素性＋統合素性	91.93	55.45
Yu 2008 [30]	87.26	N/A
Zhao 2009 [31]	87.0	N/A

比較でも、英語では発表されている最高精度のシステムと同等の精度を達成している。なお、Suzuki 2009も半教師有り学習の考え方を用いた手法であるが、Suzuki 2009の手法は、実装が我々の手法より複雑である。中国語においては、発表されている最高精度のシステムを大きく上回る精度を達成しており、我々の知る限り世界最高性能の係り受け解析器である*3。

4.3 ALAGINでの公開

ここで開発した中国語係り受け解析器は、「CNP（A ChiNese dependency Parser)」という名称でALAGIN言語資源サイトを通じてオープンソースソフトウェアとして一般公開している（http://alaginrc.nict.go.jp/cnp/index.html)。また、同時に中国語処理用のモデルパラメータを含むデータベースもALAGINから配布している。

5 まとめ

本章では、情報分析研究室で研究開発を行い、ALAGINから公開している基盤的言語処理ツール（評価情報分析器、形態素解析器、構文解析器）について解説を行った。**2**では、評価表現の抽出手法と評価表現のタイプ別分類、評価保持者の判定、および評価極性分類手法を含む評価表現分析システムについて述べた。評価表現コーパスを用いた実験を行い、システムの性能を調べた。今後の課題としては、辞書やコーパスの拡充や素性の改良によるシステムの性能を改善すること、他の言語へ拡張することが挙げられる。**3**では、パイプラインによる中国語単語分割と品詞タグ付けにおいて、簡単かつ有効な半教師あり手法を提案した。提案手法はラベルありデータを生かし、大規模なラベルなしデータから形態素情報を捉え、解析性能を向上させることができる。実験により、提案手法がベースラインおよび既存手法より高い解析精度を達成することが分かった。**4**では、係り受け解析において、大量の生コーパスをベースラインモデルで解析した結果から抽出した部分木を利用するという半教師あり学習を提案し、中国語において世界最高性能の精度を達成した。**3.5**のCSPとあわせてこれらの基礎的言語処理ツールは、現在研究室内外の様々な研究やプロジェクトで利用されている。今後もこれらのツールの精度を向上させるとともに、新たな基礎的処理の研究開発に取り組んでいきたい。

*3　論文の発表、査読時。

参考文献

1 川田拓也，中川哲治，森井律子，宮森恒，赤峯享，乾健太郎，黒橋禎夫，木俵豊，“Webテキストにおける評価情報の整理・分類およびタグ付きコーパスの構築，”言語処理学会第14回年次大会発表論文集，2008.
2 http://alaginrc.nict.go.jp/resources/nictmastar/resource-info/abstract.html#A-1
3 http://alaginrc.nict.go.jp/resources/nictmastar/resource-info/abstract.html#D-1
4 http://alaginrc.nict.go.jp/resources/nictmastar/resource-info/abstract.html#A-3
5 Eric Breck,Yejin Choi, and Claire Cardie,“Identifying expressions of opinion in context,”Proceedings-IJCAI-2007, 2007.
6 Bo Pang and Lillian Lee,“Opinion Mining and Sentiment Analysis,”Foundations and Trends in Information Retrieval, Vol. 2, No. 1–2, pp. 1–135, 2008.
7 乾孝司，奥村学，“テキストを対象とした評価情報の分析に関する研究動向，”自然言語処理，Vol. 13, No. 3, pp. 201–241, 2006.
8 Tetsuji Nakagawa, Kentaro Inui, and Sadao Kurohashi,“Dependency Tree-based Sentiment Classification using CRFs with Hidden Variables,”In Proceedings of HLT-NAACL 2010, 2010.
9 Yue Zhang and Stephen Clark,“A Fast Decoder for Joint Word Segmentation and POS Tagging Using a Single Discriminative Model,”In Proceedings of EMNLP-2010, 2010

10 Canasai Kruengkrai, Kiyotaka Uchimoto, Jun'ichi Kazama, Yiou Wang, Kentaro Torisawa, and Hitoshi Isahara, "An Error-Driven Word-Character Hybird Model for Joint Chinese Word Segmentation and POS Tagging," In Proceedings of ACL-IJCNLP-2009, 2009.

11 Canasai Kruengkrai, Kiyotaka Uchimoto, Jun'ichi Kazama, Yiou Wang, Kentaro Torisawa, and Hitoshi Isahara, "Joint Chinese Word Segmentation and POS Tagging Using an Error-Driven Word-Character Hybrid Model," IEICE transactions on information and systems 92(12), 2009.

12 Wenbin Jiang, Liang Huang, Qun Liu, and Yajuan Lu, "A Cascaded Linear Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging," In Proceedings of ACL-2008, 2008.

13 Tetsuji Nakagawa and Kiyotaka Uchimoto, "Hybrid Approach to Word Segmentation and POS Tagging," In Proceedings of ACL Demo and Poster Sessions, 2007.

14 Rie Kubota Ando and Tong Zhang, "A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data," Journal of Machine Learning Research, 2005.

15 Jun Suzuki and Hideki Isozaki, "Semi-Supervised Sequential Labeling and Segmentation using Gigaword Scale Unlabeled Data," In Proceedings of ACL-08: HLT, 2008.

16 Jun Suzuki, Hideki Isozaki, Xavier Carreras, and Michael Collins, "An Empirical Study of Semi-supervised Structured Conditional Models for Dependency Parsing," In Proceedings of EMNLP-2009, 2009.

17 Wenliang Chen, Jun'ichi Kazama, Kiyotaka Uchimoto, and Kentaro Torisawa, "Improving Dependency Parsing with Subtrees from auto-Parsed Data," In Proceedings of EMNLP-2009, 2009.

18 Terry Koo, Xavier Carreras and Michael Collins, "Simple Semi-supervised Dependency Parsing," In Proceedings of ACL-2008. 2008.

19 持橋大地，鈴木潤，藤野昭典，"条件付確率場とベイズ階層言語モデルの統合による半教師あり形態素解析，" 言語処理学会第 17 回年次大会論文集，2011.

20 Peter F. Brown, Vincent J. Della Pietra, Peter V. de Souza, Jenifer C. Lai, and Robert L.Mercer, "Class-based N-gram models of natural language," Computational Linguistics, 18 (1992), pp. 467–479, 1992.

21 Wenliang Chen, Jun'ichi Kazama, Kiyotaka Uchimoto, and Kentaro Torisawa, "Improving Dependency Parsing with Subtrees from auto-Parsed Data," In Proceedings of EMNLP 2009, 2009.

22 Wenliang Chen, Jun'ichi Kazama, Kiyotaka Uchimoto, and Kentaro Torisawa, "Exploiting Subtrees in Auto-Parsed Data to Improve Dependency Parsing," Computational Intelligence, Vol. 28, Issue 3, pp. 426–451, 2012.

23 Ryan McDonald, Koby Crammer, and Fernando Pereira, "Online large-margin training of dependency parsers," In Proceedings of ACL 2005, 2005

24 Ryan McDonald and Fernando Pereira, "Online learning of approximate dependency parsing algorithms," In Proceedings of EACL2006, 2006.

25 Xavier Carreras, "Experiments with a higher-order projective dependency parser," In Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007, 2007

26 Terry Koo, Xavier Carreras, and Michael Collins, "Simple semi-supervised dependency parsing," In Proceedings of ACL-08: HLT, 2008.

27 Joakim Nivre and Ryan McDonald, "Integrating graph-based and transition-based dependency parsers," In Proceedings of ACL-08: HLT, 2008.

28 Xavier Carreras, Michael Collins, and Terry Koo, "Tag, dynamic programming, and the perceptron for efficient, feature-rich parsing," In Proceedings of CoNLL 2008, 2008.

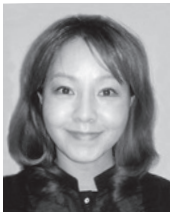
29 Jun Suzuki, Hideki Isozaki, Xavier Carreras, and Michael Collins, "An empirical study of semi-supervised structured conditional models for dependency parsing," In Proceedings of EMNLP 2009, 2009.

30 Kun Yu, Daisuke Kawahara, and Sasao Kurohashi, "Chinese dependency parsing with large scale automatically constructed case structures," In Proceedings of COLING 2008, 2008.

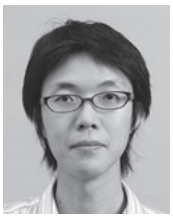
31 Hai Zhao, Yan. Song, Chunyun Kit, and Guodong Zhou, "Cross language dependency parsing using a bilingual lexicon," In Proceedings of ACL-IJCNLP 2009, 2009.

（平成 24 年 6 月 14 日 採録）

風間淳一（かざま じゅんいち）
ユニバーサルコミュニケーション研究所
情報分析研究室主任研究員
博士（情報理工学）
自然言語処理、機械学習
kazama@nict.go.jp



王　軼謳（Yiou Wang）
ユニバーサルコミュニケーション研究所
情報分析研究室研究員
博士（工学）
形態素解析、意見分析、機械翻訳、言語資源の構築
wangyiou@nict.go.jp



川田拓也（かわだ たくや）
ユニバーサルコミュニケーション研究所
情報分析研究室研究員
博士（文学）
言語学
tkawada@nict.go.jp