特集論文

自然言語処理における分野適応

Domain Adaptation in Natural Language Processing

森 信介
Shinsuke MORI
京都大学 学術情報メディアセンター
Kyoto University, Academic Center for Computing and Media Studies
forest@i.kyoto-u.ac.jp

1. はじめに

一般の人々が商品やサービスあるいは施策を評する文をウェブに書き込んでいる。日常の由無し事も非常時の切迫した状況も我々はウェブに書き連ねている。このような文章が社会的な影響力を持つにつれて、これらの機械による処理、すなわち自然言語処理[*1]への要求と期待が高まっている。また、カルテや企業の業務報告など、ウェブでは捉えられないテキストに対する処理の要求も依然として高い。

自然言語処理の研究は、電子化が早かった新聞記事や辞書の例文などを対象としてきた。その結果、これら新聞記事などの分野を中心に、処理のための情報が付与された辞書やコーパスなどの言語資源が整備された。その努力により、形態素解析や構文解析などのような基礎的な自然言語処理は、新聞記事などに対して高い解析精度を実現した。

しかしながら、現実の自然言語処理の対象は、言語資源が整備された一般分野と性質の異なる分野のテキストである。例えば、東日本大震災のときの twitter への書き込みには、twitter 特有の表現に加えて被災地の字名や方言が多数含まれていた。このような書き込みから安否情報を抽出するには、人名や場所などの固有表現を高い精度で認識することが非常に重要である。また、企業の営業担当者が書く業務報告には、その企業が取り扱う商品やサービスの名前とそれに関わる表現が頻出する。従業員に出先で業務報告を入力させるには、このような表現に対応した音声認識や仮名漢字変換システムが必要である。これらを構築するためには、分野特有の言語表現を的確に単語分割し読みを推定しての言語モデルを構築することが不可欠である。

本稿では、上述のような要求に答えるために、既存の一般分野の言語資源に少量の適応分野の言語資源を追加することで、適応分野での高い精度を実現することを目的とする分野適応について述べる。まず自然言語処理を概観し、形態素解析や構文解析の分野適応について述べる。次に、音声認識や仮名漢字変換のための言語モデルの分野適応について説明する。最後に、ある自然言語処理に必要な情報を別の自然言語処理の利用ログから引き出す研究を紹介する。

*1 本稿では、書き言葉と話し言葉の両方を自然言語と呼び、いわゆる音声言語処理の一部も自然言語処理に含める。

2. 自然言語処理の現状

人は、意思疎通や記録のために言語を用いる。これは、自然に発生したと考えられており、プログラミング言語と区別するために、自然言語と呼ばれる。このような自然言語を処理する能力を機械で実現しようというのが自然言語処理である。

自然言語処理は、入力を自然言語とする解析系と出力を自然言語とする生成系に大別できる。よく知られた解析系は、形態素解析や構文解析である。生成系の代表は、音声認識と仮名漢字変換であろう。翻訳や要約のように、入力と出力が共に自然言語である課題もある。この節では、これらの自然言語処理の課題の現状を概説する。

2·1 単語分割

単語分割は、入力文を単語に分割する処理である。単語分割に加えて、各単語の品詞と原形 (活用語の場合) も推定する処理を**形態素解析**と呼ぶ。例えば、入力文が「学校に行った」である場合の形態素解析の出力例は以下の通りである。

学校/名詞 に/助詞 行っ/動詞/行く た/助動詞

この例では、「行っ」は活用語であり、原形として「行く」が付与されて、「行う」と区別されている。

単語の定義と品詞体系 (以下では両方を指して品詞体系と呼ぶ) はいくつかあり、形態素解析システム (ツール) やコーパスによって異なる。人手で調整したコストに基づく形態素解析システム JUMAN では、システムと品詞体系が不可分である。これに対して、茶筌 [松本 96] や MeCab [工藤 04] や京都テキスト解析ツールキット (KyTea)[森 11a] は、学習に基づく方法を採用している。これらの品詞体系は、単語境界や品詞が付与された学習コーパスによって規定される。茶筌と MeCab の配布モデルは、IPA

品詞体系を採用している。KyTea の配布モデルでは、国立国語研究所の短単位 [小椋 08] に活用語尾の分割を行う改変を施している。

音声認識の言語モデルの作成や統計的機械翻訳など、単語の品詞や原形を必ずしも必要としない応用がある。このため、形態素解析を単語分割と品詞推定に分解し、多段の処理で実現する設計も考えられる。活用語の語尾を分割する場合、その原形は多くの場合に品詞から明らかなので、原形を推定しない。音声認識や音声合成などの音声言語処理や仮名漢字変換などでは、むしろ読み*2の推定が重要である。KyTea の配布モデルは、問題を単語分割と品詞推定と読み推定に分割している。なお、中国語では、形態素解析とは呼ばれず、単語分割と品詞推定と呼ばれる。KyTea には、学習コーパスを中国語にした中国語モデルも配布されている。3 章では、日本語の単語分割の分野適応について詳説する。

2·2 構 文 解 析

構文解析は、文の構造を明らかにする処理である。多くの場合、入力は品詞が付与された単語の列である。句構造文法を採用する方法 [Collins 03] [Charniak 05] と単語間の係り受けを記述する方法 [Nivre 04] [McDonald 05] [McDonald 11](**係り受け解析**とも呼ばれる) がある。日本語では、単語の代わりに文節*3を単位とすることが多いが、複合名詞の構造などのより詳細な情報付与を実現する単語係り受けも研究されている。日本語の文節単位の係り受け解析のうちの代表的なツールは、主に人手で調整したコストに基づく KNP [黒橋 95] と機械学習を用いる CaboCha [工藤 02] である。単語を単位とする係り受け解析のツールとしては、部分的アノテーションからの学習が可能な EDA [Flannery 11] がある。これらの入力は、形態素解析の結果、すなわち、品詞が付与された単語列である。

文節係り受け解析における一般分野のコーパス作成コストを低減することを目指して、品詞が付与されていない単語列からの学習や係り先が右隣の文節かそれ以外かだけをコーパス付与する場合についての報告がある [Sassano 05]。また、文内の一部の文節にのみ係り先が付与されたコーパスからの学習と能動学習のシミュレーションの実験報告がある [Sassano 10] 。単語係り受け解析では、文節係り受けコーパスを単語係り受けに変換して得られる部分的アノテーションコーパスを用いた分野適応の実験結果が報告されている [Flannery 11]。句構造文法では、部分的に付与された句構造からの確率的文脈自由文法の学習が提案されている [Pereira 92]。

∗2 正確には、音声認識の場合は発音であり、仮名漢字変換の場合は入力記号列である。両者の主な違いは、アラビアやアルファベットの列 (例: 3/3/さん) と母音の長音化 (例: 経済/けーざい/けいざい) である。

∗3 一般に、文節は、1 個以上の内容語と 0 個以上の機能語からなる単語列である。

2·3 確率的言語モデル

確率的言語モデル [北 99] は、ある言語の文の生成確率をモデル化する。統計的仮名漢字変換 [森 99][Chen 00] や音声認識 [鹿野 01] は、確率的言語モデルによる生成確率を参照して、平仮名列や発音列から尤もらしい文を出力する。確率的言語モデルは、単語や単語列の頻度に基づいている。したがって、分野適応に際しては、適応分野のテキストを用意することと単語分割や読み推定の分野適応をすることが重要である。言語モデルの分野適応については、4 章で詳説する。

2·4 そ の 他

統計的機械翻訳 [Brown 90] [Koehn 10] は、複数の自然言語処理の複合である。まず、同じ内容を原言語と目的言語で書かれた文対 (並行コーパス) を用意する。両言語に対して何らかの解析を行い、翻訳単位の対応関係を学習する。多くの方法で翻訳単位は単語であり、日本語や中国語では単語分割が必要となる。また、固有名詞などは翻字 [Knight 98] されることが多く、読み推定の結果を利用する。最近では、入力文が木構造になっていることを仮定する手法 [Lin 04] もあり、その場合には構文 (係り受け) 解析が必要となる。

統計的機械翻訳では、並行コーパスとは別に目的言語の単言語コーパスを用意し、そこから構築された言語モデルを参照する。ある分野のテキストに対して高い翻訳精度を達成するためには、単語分割や構文解析の分野適応に加えて、目的言語の言語モデルの分野適応も重要である。言語モデルの分野適応については、4 章で説明するが、統計的機械翻訳に特化した並行コーパスの分野適応 [Axelrod 11] の研究もある。

多義性解消は、複数の意味がある単語のある文脈中での意味を推定する課題である。この多義性解消の課題に対して、能動学習による分野適応を行った結果が報告されている [Chan 07]。また、固有表現認識は、製品名や組織名あるいは日時や量などの一つの実態を指す単語列を同定する課題である。固有表現の認識は、テキストマイニング [ローネン 07] などにおいて重要である。固有表現認識に対しても能動学習が試みられている [Tomanek 09]。固有表現認識の分野適応の課題は、認識すべき固有表現が応用によって異なるので、一般分野のテキストに付与された典型的な固有表現タグが必ずしも有用ではないという点であろう。

3. 単語分割の分野適応

単語分割は、日本語に対する自然言語処理のほとんどの応用で用いられる。したがって、対象となるテキストの分野 (適応分野) での単語分割の精度が重要であるが、一般分野での精度を大きく下回ることがしばしばである。しかしながら、自然言語処理をツールとして用いている

多くの研究では、辞書への単語の追加程度の対策しか取られない。こうした対策の問題点と、より多くの言語資源を用いた分野適応について説明する。

3·1 利用可能な適応分野の言語資源

自然言語処理を応用すべき課題 (例: レントゲンの読影結果の音声入力) に対して、多くの場合にその分野に関する次の 2 つの言語資源が利用可能である。

(1) 適応分野の用語集: 人のために作られた適応分野の単語リストで、ほとんどの場合に一般分野の単語分割基準には合致せず品詞も付与されていない。しばしば、読みなどの付加情報がある (例: 病名や体の部位のリスト)。

(2) 適応分野の生テキスト: 過去に蓄積された適応分野の例文集で、単語境界や品詞などの情報のない単なる文からなる (例: これまでの電子化されたレントゲンの医療所見)。

これらの言語資源を用いて適応分野の単語分割の精度を向上させることが課題である。最も単純な方法は、適応分野の用語集に含まれる見出し語を単語分割器の辞書に加えることである。ChaSen や MeCab では、単語分割が目的であっても品詞を付与する必要があるので、全ての単語を普通名詞とする。このようにして得られる単語分割器を用いると、必ずしも単語分割基準には合致しないものの、辞書に含まれる単位で単語を認識することができる。また、未知語の周辺の分割誤りも大幅に軽減できる。一方、生テキストの利用方法は自明ではない。まったく人手を介さない方法として、未知語候補を自動抽出し辞書に追加する方法が提案され、精度向上が報告されている [森 98]。茶筌などのように隠れマルコフモデルに基づいている場合には、EM アルゴリズムを用いることで生コーパスからパラメータを推定することが原理的には可能である [竹内 97]。

3·2 言語資源の追加による分野適応

上述の教師なし学習では、精度向上の程度が大きくない。したがって、絶対的な精度を重視する現場では、これらの言語資源に人手による作業を加える。

まず、3·1 節の (1) の適応分野の用語集の利用方法について述べる。辞書の見出語は、以下の 3 種類に分類できる。

- 単語 (単語分割基準に合致)
  例: |言-語|
- 複合語 (両端のみ基準に一致)
  例: |計␣算␣言␣語␣学|
- 単語列
  例: |計-算|言-語|学|

ここで、例の中の文字間の記号「|」と「-」と「␣」は、順に、単語境界が有る、無い、有るか無いか不明を表す。人用の辞書の多くの見出し語は複合語で、両端のみが信頼できる単語境界情報である。文献 [森 11b] は、複合語をそのまま用いた場合、人手の作業を加えて単語列にした場合、単語列を単語に分解して辞書に加えた場合の単語分割精度を報告している。報告によれば、複合語のままでも精度向上が見られるが、人手を加えて単語列とすることにより大きく精度が向上する。単語列を単語に分解すると、単語連接の情報が失われ、単語列として参照するよりも少し精度が低下する。この作業は、自動抽出された未知語候補に対しても同様に行うことができる。

表 1 言語資源の追加による単語分割の分野適応

言語資源	KyTea	MeCab	茶筌
辞書			
単語	○	○[1]	○[1]
複合語 (人用の辞書)	○	×	×
単語列	○	△[2]	△[2]
コーパス			
フルアノテーション	○	△[3]	△[3]
部分的アノテーション	○	×	×

○[1]: 品詞とコストの付与も必要
△[2]: フルアノテーションコーパスとして追加 (△[3])、または構成する各単語を個別に辞書に追加 (○[1])
△[3]: 実質的に不可能 (配布モデルの学習コーパスが必要)

次に、3·1 節 (2) の生テキストに関してである。適応分野の生テキストは、まず実際に解析してみて、解析精度がどの程度かを目視で推測することに用いられる。その結果、解析誤りが散見され、大部分が単語分割ツールの未知語に起因することに気付く。この誤りの対処として、未知語を単語分割ツールの辞書に追加する。多くの応用研究での分野適応は、この作業までである。未知語に起因しない誤りもあるので、単語分割精度を十分に向上させるには、生テキストへの情報付与が必須である。すなわち、文の全ての文字間またはその一部に人手で単語境界情報を付与する。こうして得られる以下の言語資源を用いて、自動単語分割ツールのモデルを再学習する。

- フルアノテーションコーパス
  例: 電-極|端-部|と|対-向|す|る
- 部分的アノテーションコーパス
  例: 電␣極|端-部|と␣対␣向␣す␣る

ここで、例の中の文字間の記号「|」と「-」と「␣」は、順に、単語境界が有る、無い、有るか無いか不明を表す。このような言語資源には文脈情報があるので、すべての部分文字列が単語となる「上端部」のような文字列を文脈に応じて単語に分割することが可能となり、単語登録のみの場合よりも精度が高くなる[*4]。

*4 現代日本語書き言葉均衡コーパスモニター版 [前川 09] において、Yahoo!知恵袋を適応分野とし、残りを一般分野とする単語分割実験において、Yahoo!知恵袋にのみ現れる単語を文脈も含めた部分的アノテーションコーパスとして追加した場合の精度 (F 値) は 97.15%で、文脈情報を削除して単なる辞書とした追加した場合の精度 (F 値) は 96.75%であった。

以上のような言語資源を実際に活用するには、単語分割ツールがそれらに対応している必要がある。表1は、主要な単語分割(形態素解析)ツールの対応状況である。形態素解析ツールのMeCabや茶筌では、単語の追加には品詞の付与が必須である。したがって、作業者は品詞体系を熟知している必要があるが、多くの現場ではそのような作業者を確保するのは困難であるので、多くの未知語は普通名詞として辞書に追加される。KyTeaでは、品詞の付与は任意であるが、モデルの再構築が必要となる。

適応分野の学習コーパスの追加は、精度向上に大きく貢献する。しかしながら、例文の全ての箇所を人手で適切に単語に分割したフルアノテーションコーパスの作成には、単語分割基準を熟知し適応分野の知識を有する作業者が必要となる。このような作業者を確保するのはほぼ不可能である。この問題に対処する方法として、KyTeaでは分野特有の表現や単語にのみ情報を付与した部分的アノテーションコーパスからの学習を可能にしている*5。学習コーパスの追加は、どのツールでもモデルの再学習が必要となる。KyTeaは、素性頻度ファイルも配布しており、あたかも配布モデルの構築に使用した学習コーパスがあるかのように追加学習が可能である。MeCabや茶筌にはこの機能がないため、配布モデルの学習コーパスが必要となり、実質的に不可能である。実用性を考えるとこのような機能は非常に重要であろう。

部分的アノテーションコーパスを作成する際のアノテーション箇所は、自動未知語抽出の結果得られる単語候補[萩原 12]の周辺や、単語分割ツールの確信度が低い箇所とする(能動学習)と効率的である。次節では、この能動学習について述べる。

3·3 能 動 学 習

適応分野の生コーパスをより積極的に活用する方法は、これにアノテーションをして学習コーパスに加えることである。より少ないアノテーションでより高い精度を実現するために、精度向上への寄与が大きいと期待される箇所をシステムに提示させる能動学習の利用が提案されている。

自動単語分割の分野適応においても能動学習の研究がある。単語分割の問題は、各文字間に単語境界があるか否かが最小の部分問題であり、これを2値分類問題として定式化し、SVMを分類器として能動学習を適用することでアノテーション箇所数を低減できる[颯々野 06]。系列予測問題としての定式化では、一般にアノテーションの最小単位は文になるので、期待される効果が大きい箇所のみをアノテーションすることができない。文献[Tomanek 09]では、確信度の低い箇所を人手でアノテーションし、残りの箇所を自動推定の結果のまま学習に用いることでこの問題に対処し、固有表現抽出の課題に対して文単位での能動学習よりも効率的であることを示している。

以上のような能動学習の多くの論文での実験は、シミュレーションである。すなわち、予めアノテーションされたデータ(プールと呼ばれる)から一定数のサンプルを取り出し、これを学習コーパスに加えてモデルを再学習し、また次のサンプルを取り出している。実際の作業を考えると、以下のような点を考慮する必要がある。

- まとまった作業時間が必要になるアノテーション箇所を1度に作業者に提示すること
- モデルの再学習にかかる時間が十分短く、作業者を待たせないこと
- アノテーション時間は判断の難易に依存し一定ではないこと
- 作業者にとって判断が難しくアノテーションできないというのも許容すること

文献[Settles 08]では、複数人に実際にアノテーション作業をしてもらい、それを観察することで得られた傾向をアノテーション箇所選択の評価関数に反映し、より現実的な状況での効率化を報告している。

文献[Neubig 11b]では、日本語の単語分割において、実際の作業者を含めた能動学習の結果を報告している。自動単語分割器はKyTeaであり、現代日本語書き言葉均衡コーパスモニター版[前川 09](以下ではBCCWJと呼ぶ)を一般分野とし、医薬品情報への分野適応を課題として、以下のアノテーション戦略を比較している。

(1) フルアノテーション: 無作為に抽出された文の単語分割結果を順に修正していく。
(2) 点アノテーション: KyTea(線形SVM)が分離平面からの距離に応じて選択した100箇所の単語境界の有無を付与する。
(3) 単語アノテーション: アノテーション箇所の選択は点アノテーションと同じであるが、それが単語内の場合はその単語の直前から直後までの文字間を、単語境界の場合は前の単語の直前から後の単語の直後までの単語境界の有無を付与する。

上記の(2)と(3)が能動学習である。図1は、横軸をアノテーションしたタグ(文字間)の数とした場合の精度の変化であり、図2は、横軸をモデルの学習も含めた作業時間とした場合の精度の変化である。ともに、グラフの立ち上がりが早い方が性能が良いことを示す。図1から、点アノテーションは、アノテーション箇所数に対して最も効率的であることが分かる。しかしながら、図2から、現実の作業では単語アノテーションの方が効率的であることがわかる。作業者にとって時間を要するのは「判断すること」である。単語分割に関しては、ある文字間の単語境界の有無の判定のために単語を認定しているので、その際に作業者の意識にのぼった情報を漏れなく付与してもらうことが重要である。

単語分割の他にも、機械学習によって実用化を迎えつ

*5 部分的アノテーションコーパスの利用は、原理的には、MeCabが用いるCRFや茶筌が用いる隠れマルコフモデルでも可能である[坪井 09][竹内 97][Dempster 77]。

表 2 単語分割の分野適応の結果 (F 値)

分野	一般	医薬品情報	特許文書	料理レシピ	twitter
テスト文の数	3,680	1,250	500	728	50
適応の方法	–	フ/点/単[1]	KWIC	KWIC	能動学習
作業時間	–	11 時間	12 時間	10 時間	90 分
適応前の精度	99.32	96.75	97.25	96.70	96.52
適応後の精度	–	98.98	97.70	97.05	97.17
4 分野全てへの適応後の精度	99.34	98.98	98.20	97.12	97.17

フ/点/単[1]: フルアノテーションと点アノテーションと単語アノテーションのすべてを含む (3·3 節参照)。

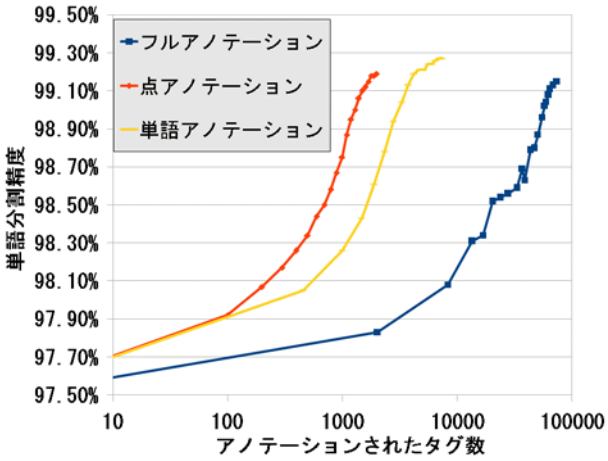


図 1 作業箇所数に対する精度向上

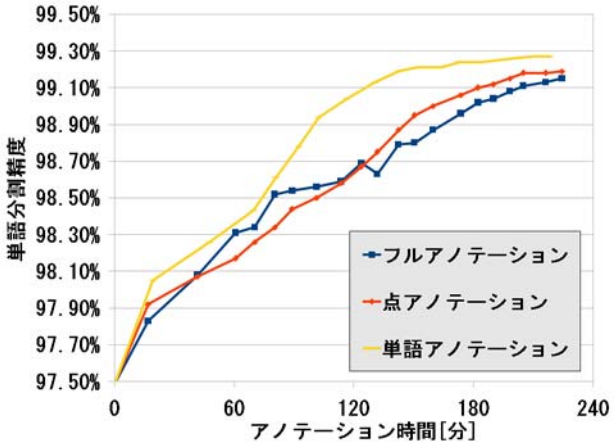


図 2 作業時間に対する精度向上

つある自然言語処理技術は多数ある。それらを実用化するには分野適応が重要であり、能動学習は非常に有用であると考えられる。その際には、アノテーションの最小単位を見極め、その単位でのアノテーションを許容するようにモデルを設計し、作業者の認知過程を考慮に入れた能動学習の枠組みを構築することが肝要である。

3·4 複数の分野適応の結果と関係

実際に単語分割の分野適応を行うと、様々な分野の部分的アノテーションコーパスが蓄積される。すると、自動単語分割のモデルは、各分野ごとに別々とするべきなのか、適応作業の結果を全て学習コーパスに加えた唯一のモデルでよいのかという問題が現れる。この問題に答えるために、BCCWJ のコアデータを一般分野とし、以下の分野適応をそれぞれ行い一般分野と適応分野での精度を測った。さらに、すべての作業結果を加えたモデルの精度を測った。

- 医薬品情報: 3·3 節で述べた分野適応実験の結果得られたコーパスをすべて利用する。
- 特許文書: NTCIR-9 [Goto 11] の特許翻訳タスクの日本語文をテストとし、NTCIR-7,8 で用いられた日本語文に対し、前後の 1 文字の参照する分布分析 (類似度計算) [森 98] を用いることで得られた未知語候補を期待頻度の降順に 3 箇所の出現箇所 (KWIC; Keyword In Context) の単語境界情報を人手で修正した。
- 料理レシピ: Web 上の料理レシピを収集し、特許文書と同様に、テスト文以外の生コーパスからの未知語候補抽出を行い期待頻度の降順に 3 箇所の出現箇所 (KWIC) を人手で修正した。
- twitter: 東日本大震災時の直後、twitter 上で特定のハッシュタグが付与された発言 [Neubig 11a] を収集し、テスト文を除いた生コーパスに対し単語アノテーションによる能動学習を行った。

表 2 は、各分野における適応作業による精度向上と、各分野の適応作業によって得られるフルアノテーションコーパスや部分的アノテーションコーパスをすべて学習データに加えたモデルによる各分野に対する精度を示している。この表の各 4 分野での適応前と適応後の精度の比較から、能動学習でも未知語候補の部分的アノテーションでも、分野適応は有効であることがわかる。さらに、最後の行の精度がいずれの分野においても最高になっていることから、別の分野への適応において得られる学習コーパスを追加しても精度が低下することはなく、場合

によっては上昇することがあることがわかる。つまり、最大の言語資源を参照する唯一のモデルを用いればよいといえる。

複数の分野の学習データを簡単に区別して用いる方法として、素性ベクトル $\boldsymbol{x}$ を拡張し、一般分野 s のデータの場合には $\boldsymbol{x}_s = (\boldsymbol{x}, \boldsymbol{x}, 0)$ とし、適応分野 t のデータの場合には $\boldsymbol{x}_t = (\boldsymbol{x}, 0, \boldsymbol{x})$ とすることが提案されている [Daume III 07]。英語の固有表現抽出と浅い構文解析での実験を報告しており、固有表現抽出において既存の複雑な手法と同等かそれ以上の精度となっている。一方、浅い構文解析では、平均的には単純に学習コーパスを加える方法と同程度の精度となっている。固有表現抽出では、ある単語列が固有表現になるか否かが分野に依存するのに対して、構文解析ではあまり依存しないことが理由と考えられる。

4. 言語モデルの分野適応

生成系の自然言語処理の代表は音声認識 [鹿野 01] と統計的仮名漢字変換 [森 99] であろう。統計的仮名漢字変換は、確実な発音と音響モデルによる音声認識といえる。音声認識では語彙を限定し、語彙以外の単語を出力しない点が主な違いである。この節では、まず音声認識について概説し、言語モデルの分野適応について述べる。なお、音響モデルの話者適応に関しては、[篠田 12] を参照されたい。

4·1 音 声 認 識

音声認識は、音響特徴量の列 $\boldsymbol{s}$ を入力とし、語彙 $\mathcal{W}_k$ の正閉包 (長さ 1 以上の任意の単語列の集合) のうち、以下の式の確率が最大となる要素 (単語列)$\hat{\boldsymbol{w}}$ を出力する。

$$\begin{aligned}\hat{\boldsymbol{w}} &= \mathop{\mathbf{argmax}}_{\boldsymbol{w}\in\mathcal{W}_k^+} P(\boldsymbol{w}|\boldsymbol{s}) \\ &= \mathop{\mathbf{argmax}}_{\boldsymbol{w}\in\mathcal{W}_k^+} \frac{P(\boldsymbol{s}|\boldsymbol{w})P(\boldsymbol{w})}{P(\boldsymbol{s})} \\ &= \mathop{\mathbf{argmax}}_{\boldsymbol{w}\in\mathcal{W}_k^+} P(\boldsymbol{s}|\boldsymbol{w})P(\boldsymbol{w})\end{aligned}$$

この式における $P(\boldsymbol{w})$ が確率的言語モデルである。

多くの確率的言語モデルは、文頭から順に単語を 1 つずつ予測する。すなわち、i 番目の単語を w_i とすると、以下の式が示すように、それを予測するときに履歴を $H_i = \boldsymbol{w}_1^{i-1} = w_1 w_2 \cdots w_{i-1}$ とする。

$$P(\boldsymbol{w}) = \prod_{i=1}^{h+1} P(w_i|H_i) \qquad (1)$$

ここで、h は文長 (単語数) であり、w_{h+1} は文末を表す特殊な記号である。よく用いられる言語モデルは、履歴を直前の $n-1$ 単語とする単語 n-gram モデルである。

$$P(w_i|H_i) = P(w_i|\boldsymbol{w}_{i-n+1}^{i-1}) \qquad (2)$$

この確率は、単語に分割されたコーパスから以下の式を用いて最尤推定される。

$$P(w_i|\boldsymbol{w}_{i-n+1}^{i-1}) = \frac{f(\boldsymbol{w}_{i-n+1}^{i})}{f(\boldsymbol{w}_{i-n+1}^{i-1})} \qquad (3)$$

ここで $f(\boldsymbol{w})$ は、コーパスにおける単語列 $\boldsymbol{w}$ の頻度である。日本語などの単語境界が明示されない言語に対しては、適応分野のコーパスを自動で単語分割することになる。したがって、前章で述べたような自動単語分割器の適応をすることが望ましい[*6]。

4·2 言語モデルの分野適応

式 (3) の確率値が正確であるために、音声認識の対象とする分野の文の分布を反映する大量の文から推定することが望ましい。しかしながら、新聞や Web などの文を認識対象とする場合を除けば、これらに比肩するほどの量の適応分野の文が利用可能であることはまれである。例えば、医療所見や業務報告の音声入力システムを作成する場合には、それまでに蓄積した医療所見や業務報告を用いることになる。しかし、このような適応分野のコーパスは十分大きくない場合が多い。このような場合には、一般分野の言語モデルを対象の分野に適応する。この目的でよく用いられる方法は以下の式で表わされる補間である。

$$P(w_i|H_i) = \lambda_g P_g(w_i|H_i) + \lambda_t P_t(w_i|H_i) \qquad (4)$$

この式中の P_g と P_t はそれぞれ、一般分野の単語分割済みコーパス C_g から推定した確率と適応分野の単語分割済みコーパス C_t から推定した確率を表す。さらに λ_g と λ_t は両モデルの補間係数であり、$\lambda_g + \lambda_t = 1$ である。これらは、例えば以下の削除補間法 [Jelinek 91] により推定する。

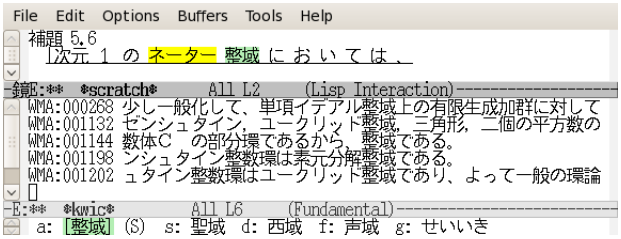
(1) 適用分野のコーパスを k 個に分割し $C_{t,j}$ ($1 \le j \le k$) を得る。
(2) 各 j に対し、$C_{t,j}$ を除いた $k-1$ 個の部分コーパスから言語モデル $P_{t,j}(w_i|H_i)$ を推定する。
(3) 言語モデル $\lambda_g P_g(w_i|H_i) + \lambda_t P_{t,j}(w_i|H_i)$ によるコーパス $C_{t,j}$ の出現確率の j に対する幾何平均が最大になるように λ_g と λ_t を決定する。

この手続きで、モデル推定のコーパスと最適化の対象のコーパスを別にしているのは、適応分野の未知のテストデータを模擬するためである。これをしないと、λ_t が過剰に高くなる。

より簡便な方法として、以下の式のように、適応分野のコーパスの頻度に一定の重み α を掛けて一般分野のコーパスの頻度と加算して確率を推定することもある。

$$P(w_i|\boldsymbol{w}_{i-n+1}^{i-1}) = \frac{f_g(H_i, w_i) + \alpha f_t(H_i, w_i)}{f_g(H_i) + \alpha f_t(H_i)}$$

*6 音声認識と仮名漢字変換のいずれにおいても、読み推定の分野適応はより重要である。ただし、多くの場合、適切に単語に分割されていれば、辞書を充実するだけで十分である。



数学用語「整域」が Wikipedia の数学関連のページの部分文字列として読み「せいいき」の変換候補として挙げられている。画面の下半分は「聖域」の KWIC である。

図 3 変換候補として部分文字列を列挙している例

ここで、f_g と f_t はそれぞれ、適応分野のコーパスの頻度と一般分野のコーパスの頻度である。これは、式 (4) において $\lambda_g = \frac{f_g(H_i)}{f_g(H_i)+\alpha f_t(H_i)}$, $\lambda_t = \frac{\alpha f_t(H_i)}{f_g(H_i)+\alpha f_t(H_i)}$ とした場合と同じである。パラメータ α は、適応分野の開発データの尤度が最大になるように決定する。

5. 自然言語処理システムの利用ログの活用

前節までで述べた自然言語処理システムの分野適応は、主にコストをかけて人手で言語資源を作成することを前提としている。この節では、自然言語処理に有用な情報を、人間の日々の言語活動から得る取り組みについて紹介する。

5·1 音声とテキストからの読みの獲得

音声認識や仮名漢字変換の言語モデル、あるいは音声合成のフロントエンドは、単語の読みを必要とする。これを実際の音声から学習する試みがある。文献 [Badr 11] は、音声とその書き起こしから単語の実際の発音を推定することで、音声認識の精度向上を実現している。しかしながら、書き起こしはコストが高いので、音声とそれに関連するテキストから未知語とその読みを抽出する方法も提案されている。文献 [Kurata 07] では、講議音声とそのテキストを用いて自動的に語彙拡張を行い、音声認識の精度向上を実現している。他に、ニュース音声とニュース記事から未知語の候補とその読みを獲得し、仮名漢字変換や音声合成のフロントエンド (言語処理部) の精度向上を実現する研究がある [笹田 10][Sasada 08]。

5·2 仮名漢字変換のログの活用

前節で述べた仮名漢字変換は、ユーザーが入力したい単語列の読みを入力し、意図した表記の単語を選択する。生テキストの全ての部分文字列も変換候補として列挙することができる仮名漢字変換システム [森 07] を用いると、仮名漢字変換のログにユーザーの意図した単語とその読みが記録される。例えば、図 3 が示すように、読みが「せいいき」となる可能性があり[*7]、さらに文脈に合致する文字列として「整域」が Wikipedia の数学関連のページから挙げらる。これをユーザーが選択すると、単語「整域」が読みや文脈を伴なって獲得される。獲得された単語を単語分割と読み推定に用いる試み [森 10] や音声認識に用いる研究 [山口 12] がある。

仮名漢字変換のログは、誤確定などの誤りを多数含む。したがって、[森 10] や [山口 12] のように、単純に学習データとして用いるのではなく、より洗練された機械学習を用いることでより効率的に活用できると考えられる。

6. お わ り に

本稿では、単語分割と言語モデルを中心に自然言語処理の分野適応について述べた。分野適応は、一般分野で実用水準に達している処理に対して求められる技術である。しかしながら、実用を意図したシステムであれば、将来の分野適応を意識して設計しておくことが重要である。その際には、単語分割や言語モデルの分野適応の知見が活かされると考えられる。

分野適応技術により、学習データの作成コストは小さくなる。さらに、これをなくすことも重要であると考える。幸いにして自然言語処理に必要な情報は、人々の日常の言語活動から抽出できるはずである。人々に使ってもらえる水準のアプリケーションを作成し、その利用ログを収集できれば、誤りを含むデータからの学習 [鹿島 12] を用いて利用することが可能であろう。

謝　　辞

本論文の執筆に貢献して下さった NEUBIG Graham 博士と笹田鉄郎氏に心から感謝いたします。

◇ 参 考 文 献 ◇

[Axelrod 11] Axelrod, A., He, X., and Gao, J.: Domain Adaptation via Pseudo In-Domain Data Selection, in *Conference on Empirical Methods in Natural Language Processing*, pp. 355–362 (2011)

[Badr 11] Badr, I., McGraw, I., and Glass, J.: Pronunciation Learning from Continuous Speech, in *Proceedings of the InterSpeech2011*, pp. 549–552 (2011)

[Brown 90] Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S.: A Statistical Approach to Machine Translation, *Computational Linguistics*, Vol. 16, No. 2, pp. 79–85 (1990)

[Chan 07] Chan, Y. S. and Ng, H. T.: Domain Adaptation with Active Learning for Word Sense Disambiguation, in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pp. 49–56 (2007)

[Charniak 05] Charniak, E. and Johnson, M.: Coarse-to-fine N-best Parsing and MaxEnt Discriminative Reranking, in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 173–180 (2005)

[Chen 00] Chen, Z. and Lee, K.-F.: A New Statistical Approach To Chinese Pinyin Input, in *Proceedings of the 38th Annual Meeting of*

*7 単漢字辞書に各単語の可能な読みが列挙されている。

the Association for Computational Linguistics, pp. 241–247 (2000)
[Collins 03] Collins, M.: Head-Driven Statistical Models for Natural Language Parsing, *Computational Linguistics*, Vol. 29, No. 4, pp. 589–637 (2003)
[Daume III 07] Daume III, H.: Frustratingly Easy Domain Adaptation, in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pp. 256–263 (2007), Companion Volume Proceedings of the Demo and Poster Sessions
[Dempster 77] Dempster, A. P., Laird, N. M., and Rubin, D. B.: Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, Vol. 39, No. 1, pp. 1–38 (1977)
[Flannery 11] Flannery, D., Miyao, Y., Neubig, G., and Mori, S.: Training Dependency Parsers from Partially Annotated Corpora, in *Proceedings of the Fifth International Joint Conference on Natural Language Processing* (2011)
[Goto 11] Goto, I., Lu, B., Chow, K. P., Sumita, E., and Tsou, B. K.: Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop, in *Proceedings of NTCIR-9 Workshop Meeting*, pp. 559–578 (2011)
[Jelinek 91] Jelinek, F., Mercer, R. L., and Roukos, S.: Principles of Lexical Language Modeling for Speech Recognition, in *Advances in Speech Signal Processing*, chapter 21, pp. 651–699, Dekker (1991)
[Knight 98] Knight, K. and Graehl, J.: Machine Transliteration, *Computational Linguistics*, Vol. 24, pp. 599–612 (1998)
[Koehn 10] Koehn, P.: *Statistical Machine Translation*, Cambridge University Press (2010)
[Kurata 07] Kurata, G., Mori, S., Itoh, N., and Nishimura, M.: Unsupervised Lexicon Acquisition from Speech and Text, in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 421–424 (2007)
[Lin 04] Lin, D.: A Path-based Transfer Model for Machine Translation, in *Proceedings of the 20th International Conference on Computational Linguistics*, pp. 625–630 (2004)
[McDonald 05] McDonald, R., Pereira, F., Ribarov, K., and Hajič, J.: Non-projective Dependency Parsing Using Spanning Tree Algorithms, in *Conference on Empirical Methods in Natural Language Processing*, pp. 523–530 (2005)
[McDonald 11] McDonald, R. and Nivre, J.: Analyzing and Integrating Dependency Parsers, *Computational Linguistics*, Vol. 37, No. 4, pp. 197–230 (2011)
[Neubig 11a] Neubig, G., Matsubayashi, Y., Hagiwara, M., and Murakami, K.: Safety Information Mining - What can NLP do in a disaster -, in *Proceedings of the Fifth International Joint Conference on Natural Language Processing* (2011)
[Neubig 11b] Neubig, G., Nakata, Y., and Mori, S.: Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (2011)
[Nivre 04] Nivre, J. and Scholz, M.: Deterministic Dependency Parsing of English Text, in *Proceedings of the 20th International Conference on Computational Linguistics*, pp. 64–70 (2004)
[Pereira 92] Pereira, F. and Schabes, Y.: Inside-Outside Reestimation from Partially Bracketed Corpora, in *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pp. 128–135 (1992)
[Sasada 08] Sasada, T., Mori, S., and Kawahara, T.: Extracting Word-Pronunciation Pairs from Comparable Set of Text and Speech, in *Proceedings of the InterSpeech2008*, pp. 1821–1824 (2008)
[Sassano 05] Sassano, M.: Using a Partially Annotated Corpus to Build a Dependency Parser for Japanese, in *Proceedings of the Second International Joint Conference on Natural Language Processing*, pp. 82–92 (2005)
[Sassano 10] Sassano, M. and Kurohashi, S.: Using Smaller Constituents Rather Than Sentences in Active Learning for Japanese Dependency Parsing, in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 356–365 (2010)
[Settles 08] Settles, B., Craven, M., and Friedland, L.: Active Learning with Real Annotation Costs, in *NIPS Workshop on Cost-Sensitive Learning* (2008)
[Tomanek 09] Tomanek, K. and Hahn, U.: Semi-Supervised Active Learning for Sequence Labeling, in *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, pp. 1039–1047 (2009)
[ローネン 07] ローネン フェルドマン, ジェイムズ サンガー：テキストマイニングハンドブック, 東京電機大学出版局 (2007)
[工藤 02] 工藤 拓, 松本 裕治:チャンキングの段階適用による日本語係り受け解析, 情報処理学会論文誌, Vol. 43, No. 6, pp. 4834–1842 (2002)
[工藤 04] 工藤 拓, 山本 薫, 松本 裕治：Conditional Random Fields を用いた日本語形態素解析, 情報処理学会研究報告, 第 NL161 巻 (2004)
[黒橋 95] 黒橋 禎夫, 長尾 眞：並列構造の検出に基づく長い日本語文の構文解析, 自然言語処理, Vol. 1, No. 1, pp. 35–57 (1995)
[笹田 10] 笹田 鉄郎, 森 信介, 河原 達也：自動獲得した未知語の読み・文脈情報による仮名漢字変換, 自然言語処理, Vol. 17, No. 4, pp. 131–154 (2010)
[山口 12] 山口 洋平, 森 信介, 河原 達也：仮名漢字変換ログを用いた講義音声認識のための言語モデル適応, 言語処理学会第 18 回年次大会発表論文集 (2012)
[鹿島 12] 鹿島 久嗣, 梶野 洸：クラウドソーシングと機械学習, 人工知能学会誌, Vol. 27, No. 4 (2012)
[鹿野 01] 鹿野 清宏, 伊藤 克亘, 河原 達也, 武田 一哉, 山本 幹雄：音声認識システム, オーム社 (2001)
[篠田 12] 篠田 浩一：音声認識における転移学習: 話者適応, 人工知能学会誌, Vol. 27, No. 4 (2012)
[小椋 08] 小椋 秀樹, 小磯 花絵, 冨士池 優美, 原 裕：『現代日本語書き言葉均衡コーパス』形態論情報規程集, 独立行政法人国立国語研究所 (2008)
[松本 96] 松本 裕治：形態素解析システム「茶筌」, 情報処理, Vol. 41, No. 11, pp. 1208–1214 (1996)
[森 98] 森 信介, 長尾 眞：n グラム統計によるコーパスからの未知語抽出, 情報処理学会論文誌, Vol. 39, No. 7, pp. 2093–2100 (1998)
[森 99] 森 信介, 土屋 雅稔, 山地 治, 長尾 真：確率的モデルによる仮名漢字変換, 情報処理学会論文誌, Vol. 40, No. 7, pp. 2946–2953 (1999)
[森 07] 森 信介：無限語彙の仮名漢字変換, 情報処理学会論文誌, Vol. 48, pp. 3532–3540 (2007)
[森 10] 森 信介, Neubig, G.：仮名漢字変換ログの活用による言語処理精度の自動向上, 言語処理学会年次大会 (2010)
[森 11a] 森 信介, Graham, N., 坪井 祐太：点予測による単語分割, 情報処理学会論文誌, Vol. 52, No. 10, pp. 2944–2952 (2011)
[森 11b] 森 信介, 小田 裕樹：3 種類の辞書による自動単語分割の精度向上, 自然言語処理, Vol. 18, No. 2 (2011)
[前川 09] 前川 喜久雄：代表性を有する大規模日本語書き言葉コーパスの構築, 人工知能学会誌, Vol. 24, No. 5, pp. 616–622 (2009)
[竹内 97] 竹内 孔一, 松本 裕二：隠れマルコフモデルによる日本語形態素解析のパラメータ推定, 情報処理学会論文誌, Vol. 38, No. 3, pp. 500–509 (1997)
[坪井 09] 坪井 祐太, 森 信介, 鹿島 久嗣, 小田 裕樹, 松本 裕治：日本語単語分割の分野適応のための部分的アノテーションを用いた条件付き確率場の学習, 情報処理学会論文誌, Vol. 50, No. 6, pp. 1622–1635 (2009)
[萩原 12] 萩原 正人, 関根 聡：半教師あり学習に基づく大規模語彙に対応した日本語単語分割, 言語処理学会第 18 回年次大会発表論文集 (2012)
[北 99] 北 研二：確率的言語モデル, 言語と計算 (4), 東京大学出版会 (1999)
[颯々野 06] 颯々野 学：日本語単語分割を題材としたサポートベクタマシンの能動学習の実験的研究, 自然言語処理, Vol. 13, No. 2, pp. 27–41 (2006)

〔担当委員：××○○〕

19YY 年 MM 月 DD 日　受理

著 者 紹 介

森　　信介

1998 年京都大学大学院工学研究科電子通信工学専攻博士後期課程修了. 同年日本アイ・ビー・エム (株) 入社. 2007 年より京都大学学術情報メディアセンター准教授. 京都大学博士 (工学). 1997 年情報処理学会山下記念研究賞受賞. 2010 年情報処理学会論文賞受賞. 2010 年第 58 回電気科学技術奨励賞. 言語処理学会, 情報処理学会各会員.