素性頻度ファイルと部分的アノテーションコーパスからの単語分割器の学習

笹田 鉄郎[1]　森 信介[1]　Graham NEUBIG[2]　河原 達也[1]

概要：本論文では、素性頻度ファイルと部分的アノテーションコーパスを用いて単語分割器を学習する枠組みについて提案する。一般分野のコーパスから作成した素性頻度ファイルを参照すると、そのコーパスを直接参照しているかのように単語分割器を再学習することが可能である。また、部分的アノテーションコーパスの作成により、低い人的コストで高い分野適応性を実現できる。提案する枠組みを用いて分野適応を行った結果、単語分割の精度が改善されることを確認した。

キーワード：自動単語分割，分野適応，能動学習

Training a Word Segmenter from a Feature Frequency File and Partially Annotated Corpora

Tetsuro SASADA[1]　Shinsuke MORI[1]　Graham NEUBIG[2]　Tatsuya KAWAHARA[1]

Abstract: This paper propose a framework of training a word segmenter from a feature frequency file and partially annotated corpora. A feature frequency file enable users to rebuild a word segmenter as if they use the original corpora. Patially annotated corpora make it possible to achieve domain adaptation with a minimum amount of annotation. In a domain adaptation experiment, we observed an improvement in the word segmentation accuracy.

Keywords: Word segmentation, Domain adaptation, Active learning

1. はじめに

近年では、計算機の普及に伴い、多様な分野を対象とした自然言語処理システムの研究・開発が行われている。自然言語処理の研究では、新聞記事や定型的な例文が一般分野の言語資源として用いられる。単語分割や形態素解析のような、自然言語処理の基礎となるシステムでは、一般分野の学習コーパスを増加させたり、辞書の整備を行うことによって性能を向上させてきた。

一方、自然言語処理の応用における対象は、言語資源が整備された一般分野とは異なる分野であることが多い。ここでまず最初に問題となるのは対象となるテキストの単語分割精度であり、分割誤りを残したまま後段の処理を行うと応用全体の性能に悪影響を及ぼす。このため、自然言語処理の応用においては対象のテキストに応じた分野適応を行い、単語分割器の再学習を行うことが不可欠である。

通常、単語分割器の再学習は、一般分野のコーパスと適応分野のコーパスを用意して行う。一般分野のコーパスは大規模なものを用いることが望ましいが、そのようなコーパスをツールの作成者が再配布することは実質的に不可能であり、ユーザが個々にコーパスを手に入れなければならないという問題が起こる。また、分野適応をタスクに応じて効率よく行うことも必要となる。

本論文では上述の課題を解決するために、テキスト解析

[1] 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
[2] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology

器 KyTea [*1] を用い、素性頻度ファイルと部分的アノテーションコーパスによる分野適応を行って単語分割器の再学習を行う枠組みを提案する。提案する枠組みにおいては、個々のユーザは一般分野の大規模コーパスを用いて学習する場合と同様の条件で単語分割器の再学習を行うことが可能となる。また、少量の適応分野の言語資源に対して部分的アノテーションコーパスを作成することにより、効率的な分野適応を行うことができる。実験では、特許文書、料理レシピ、twitter の 3 分野を対象として分野適応を行い、単語分割の精度が改善することを確認した。

提案する枠組みで構築した単語分割器で評価実験を行った結果、部分的アノテーションによる分野適応を用いることで効率的に精度が向上することを確認した。

2. 素性頻度ファイルを用いた単語分割器の学習

単語分割は、入力文を単語に分割する処理である。1 節で述べたように、単語分割システムを分野適応のためのツールとして用いるユーザは、それぞれのタスクに応じて適切な単語分割器を構築可能であることが望ましい。本節では点予測による単語分割器で用いる素性と、その結果得られる素性頻度ファイルを用いた分野適応について述べる。

点予測による単語分割 [1] では、入力を文字列 $\boldsymbol{x} = x_1x_2\cdots x_n$ として、各文字間に単語境界の有無を示すタグ $\boldsymbol{t} = t_1t_2\cdots t_{n-1}$ を出力する。単語境界タグ t_i がとりうる値は、文字 x_i と x_{i+1} の間に単語境界が「存在する」か「存在しない」の 2 種類で、2 値分類問題として定式化される。点予測による単語分割では、以下の 3 種類の素性を参照する SVM による分類を行っている。

(1) 文字 n-gram: 単語境界の判定対象となるタグ位置 i の前後の部分文字列であり、窓幅 m と長さ n のパラメータがある。素性は、長さ $2m$ の文字列 $x_{i-m+1},\cdots,x_{i-1},x_i,x_{i+1},\cdots,x_{i+m}$ の長さ n 以下のすべての部分文字列 (文字 n-gram) である。

(2) 文字種 n-gram: 文字を文字種に変換した列を対象とする点以外は文字 n-gram と同じである。文字種は、漢字（K)、片仮名（k)、平仮名（H)、ローマ字（R)、数字（N)、その他（O）の 6 つである。

(3) 単語辞書素性: タグ位置 i を始点とする単語、終点とする単語、内包する単語が辞書にあるか否かのフラグと、その単語の長さである。

テキスト解析器 KyTea には、上述の素性と単語境界の有無を示すラベルの組を素性頻度ファイルとして出力する機能が追加されている。一度素性頻度ファイルを出力しておけば、目的に応じてコーパスを追加することで、元のコーパスと追加したコーパスの両方を参照している場合と全く同じように単語分割器を再学習することが可能である。

また、KyTea の配布元では、『現代日本語書き言葉均衡コーパス』正式公開版 [2] (以下、BCCWJ) のコアデータならびに UniDic [3] などから上述の素性について頻度を計数した素性頻度ファイルを配布しており[*2]、大規模な一般分野のコーパスを所持していないユーザでも、タスクに応じた単語分割器の分野適応を容易に行うことができる。

3. 単語分割器の分野適応

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:レシピコーパスを用いた調理手順の可視化) 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: 料理の名前や材料のリスト)。

(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: 調理の手順)。

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。形態素解析ツールの MeCab や茶筌では、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 [4]。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 [5]。

[*1] http://www.phontron.com/kytea/index-ja.html

[*2] http://www.phontron.com/kytea/train-ja.html#feature

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	○	○[1]	△[2]
コーパス			
フルアノテーション	○	○[1]	△[3]
部分的アノテーション	○	×	×

○[1] : 品詞の付与も必要
△[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)

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。特に、3.1 項 (2) のような適応分野の生テキストに対し、何らかの形でアノテーションを行い、学習コーパスとして追加することが多い。

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[*3]。

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。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[*4]。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。また、2 節で述べたとおり、KyTea の配布元では素性頻度ファイルも配布しているため、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。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 [9] 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 [10]。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 [11] 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

[*3] 現代日本語書き言葉均衡コーパスモニター版 [6] 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。

[*4] 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 [7][5][8]。

表 2 コーパス

出典	用途	文数	単語数	文字数
BCCWJ	学習	53,899	1,275,135	1,834,874
特許文書	学習	2,322	—	171,090
料理レシピ	学習	1,651	—	85,881
twitter	学習	324	—	25,668
BCCWJ	テスト	6,406	139,229	203,541
特許文書	テスト	500	20,659	32,139
料理レシピ	テスト	728	13,248	19,967
twitter	テスト	50	2,134	3,115

表 3 単語分割の分野適応の結果 (F 値)

分野	一般	特許文書	料理レシピ	twitter
テスト文の数	6,406	500	728	50
適応の方法	–	KWIC	KWIC	能動学習
アノテーションタグ数	–	8,181	7,402	1,533
作業時間	–	12 時間	10 時間	1.5 時間
適応前の精度	98.77	94.17	96.45	94.75
適応後の精度	–	94.63	96.79	95.73
3 分野全てへの適応後の精度	98.80	94.84	96.77	95.59

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 [12] 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 [1] 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、BCCWJ を一般分野とし、医薬品情報への分野適応を課題として、単語アノテーションの有用性について述べている。具体的には、KyTea (線形 SVM) が分離平面からの距離に応じて選択した 100 箇所の単語境界を含む単語に関して、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) が能動学習であり、文献 [1] では単語アノテーションが最も有用であると報告している。

4. 評価実験

本節では、3 節で述べた単語分割器の分野適応を行い、単語分割の性能を検証する。

4.1 実験の条件

本実験では、BCCWJ のコアデータを一般分野として、特許文書、料理レシピ、twitter の 3 分野に対して、それぞれ以下に示す方法で分野適応を行い、単語分割の精度を測った。単語分割器の学習コーパスと精度評価に用いたテストコーパスを表 2 に示す。

- 特許文書: NTCIR-9 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) [4] 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 [13] 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。

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、3 分野すべての作業結果を加えたモデルの精度を測った。

4.2 複数の分野適応の結果と関係

表 3 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 3 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各分野ごとの適応における精度と比較して上昇、もしくは精度にほとんど変化がないことがわかる。なお、twitter の精度に関しては 0.14 ポイントの精度低下が見られるが、これはテストセットが他の 2 分野に比較して少ないためであり、実際に単語分割結果を比較したところ、誤りの増加している単語境界は 3 箇所のみであった。以上の結果より、分野適応を行って単語分割器を用いる場合には、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。

5. おわりに

本論文では、素性頻度ファイルと部分的アノテーションコーパスを用いた単語分割器の分野適応について述べた。提案した枠組みによって単語分割器の分野適応を行った結果、効率的に精度が改善されることを確認した。精度が要求される応用を扱う上では、分野適応にコストをかけることは避けられないため、将来誰がどのようにして作業を行うかを意識してシステムを設計することが重要である。

参考文献

[1] Neubig, G., Nakata, Y. and Mori, S.: Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (2011).
[2] 小椋秀樹，小磯花絵，冨士池優美，宮内佐夜香，小西光，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 第 4 版 (上)(下)，大学共同利用機関法人人間文化研究機構 国立国語研究所 (2011).
[3] 伝　康晴，小木曽智信，小椋秀樹，山田　篤，峯松信明，内元清貴，小磯花絵：コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用，日本語科学，Vol. 22, pp. 101–122 (2007).
[4] 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).
[5] 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).
[6] 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).
[7] 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).
[8] 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).
[9] 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).
[10] 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).
[11] Tomanek, K. and Hahn, U.: Semi-Supervised Active Learning for Sequence Labeling, *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, pp. 1039–1047 (2009).
[12] Settles, B., Craven, M. and Friedland, L.: Active Learning with Real Annotation Costs, *NIPS Workshop on Cost-Sensitive Learning* (2008).
[13] Neubig, G., Matsubayashi, Y., Hagiwara, M. and Murakami, K.: Safety Information Mining - What can NLP do in a disaster -, *Proceedings of the Fifth International Joint Conference on Natural Language Processing* (2011).