Dependent Dirichlet Process を用いた日本語文書へのタグづけのオンライン学習

熊野 正　田中 英輝
NHK 放送技術研究所
{kumano.t-eq, tanaka.h-ja}@nhk.or.jp

1　はじめに

NHK では 2012 年より、子供や日本に住む外国人を対象とした、やさしい日本語によるニュースサービス「NEWS WEB EASY (NWE)」を Web で提供している[1]。このサービスで配信しているニュースは、通常のスタイルで書かれたニュース原稿[2] のいくつかを選び、専門の作業者がやさしい日本語のスタイルに書き直すことで作成している [5]。

NWE Web ページでは読者の読解を支援するため、やさしい日本語ニュースに対して、ルビや固有名詞種別、難しい単語の語釈を付加している。これらの情報は、原稿を自動処理した結果を人手で修正することで付加している。また、書き直し作業の過程で作業者は執筆中の原稿の難しさが適切であるかどうかを随時確認しているが、これは自動処理で付与された各語の難易度を参照することで行っている。このように、NWE サービスの提供には原稿の各単語に各種情報（「タグ」）を自動付与する技術が大きく関わっており、この精度を向上させることは作業効率向上のために重要である。

これまで我々は、前述の自動タグ付与を行うために、入力を形態素解析し、各形態素に固定的にタグを付与する仕組みを提供してきた。しかし、新語は日々発生するため、形態素解析辞書や形態素タグ知識のメンテナンスが煩雑であり、また、複数形態素に 1 つのタグを付与したい場合や、同じ形態素に文脈に応じて異なるタグを付与したい場合など、原理的に対応が困難なことも多かった。

以上の問題を包括的に解決するために、本稿では、オンライン学習を用いたタグづけ手法を提案する。本手法は、新たに作成されたタグつき文書を逐次学習し、その結果を次のタグ自動付与処理に反映する。また、必要に応じて複数タグ列をひとまとまりで記憶することで、近傍文脈に応じた多義語に対するタグの使い分けを可能とする。さらに、「ある語のタグづけ方針が変更されることがある」ことを踏まえ、学習したタグづけ知識を書き換える機構を備えることとした。

[1] http://www3.nhk.or.jp/news/easy/
[2] NHK NEWS WEB (http://www3.nhk.or.jp/news/)

このような機構を実現するために、我々は dependent Dirichlet process [3] に基づく「タグつき句」unigram 生起モデルを用いている。評価実験の結果、十分な学習が行われた後の平均タグづけ正解率は 95% 以上であり、実用レベルの性能を達成することができた。

2　語彙への情報付与とその用途

NWE 作業者は、やさしい日本語原稿を複数の単位に分割し、各単位に以下の種類の情報を付与している。

- 漢字の部分の読み
- 語彙の難易度、または固有表現に対してその種別
  難易度:　級外、1〜4 級[3]、基本語
  固有表現種別:
  「人名」「地名」「組織名」「その他の固有名詞」
- 語釈: 必要に応じて、語彙を理解するのに適切な市販の小学生向け国語辞典の語釈を付与

情報付与単位は一般には単語であるが、連語など複数の単語から構成されることもあり、自明ではない。

やさしい日本語ニュース執筆においては、語釈が与えられていない難語（難易度 1〜2 級および級外）は原則として使用ないよう取り決めている。また語釈を与えた場合であっても難語の使用率を小さくするよう努めている。一方、Web でニュースを提供する際には、付与情報に基づいて、漢字のルビ、固有名詞種別（色分け）、難語の語釈を提示している。

3　「タグつき句」生起モデルを用いた「タグづけ」

「タグづけ」処理とは、例えば入力「レースの洋服です。」に対して、以下のようにその分割を決定し、かつ分割された各単位に各種情報（「タグ」）を付与するタスクである。

タグ	読み 難易度/種別 語釈	1 級 薄い布	基本	ようふく 4 級	基本	基本
表層		レース	の	洋服	です	。

[3] （旧）日本語能力試験出題基準 [2] に準拠

タグつきテキストが与えられたとき、これを学習データとするタグづけ処理を実現する簡便な方法は、⟨表層, 付与タグ⟩ の同時生起モデルを学習する手法である。我々はこれを拡張し、複数のタグからなるタグ列をより長い表層と対応づけた、⟨表層, 付与タグ列⟩（以後「タグつき句」と呼ぶこととする）の同時生起確率を学習対象とする、タグつき句 unigram 生起モデルを用いることにした。前述の例の場合、このタグつきテキストの生起が、一例として以下のような 2 つのタグつき句の列として説明できるものと考える。

	タグつき句$_1$			タグつき句$_2$	
タグ列	1 級 薄い布	基本	ようふく 4 級	基本	基本
表層	レース	の	洋服	です	。

タグつき句生起モデルを用いることにより、例えば「レース」という語に通常 {, 1 級, 競走} というタグが付与されるような場合においても、上記例の文脈における「レース」に適切なタグを付与することが可能となる。ただし、タグつき句への分割は学習データに示されていないので、学習データをどのように分割して記憶するのが適切であるかは推定する必要がある。これは教師なし形態素解析の学習問題と同等である [4]。

タグつき句への分割の推定方法　表層 w と付与タグ列 T の組であるタグつき句 $v = \langle w, T\rangle$ の unigram 生起確率が $P(v)$ と与えられているとき、m 個に分割された表層 $\boldsymbol{W}_{1:m} = w_1, \ldots, w_m$、およびその各々に付与されたタグ $\boldsymbol{T}_{1:m} = t_1, \ldots, t_m$ の組からからなるタグつきテキスト $\boldsymbol{U}_{1:m}$ がどのようなタグつき句列として生起したと考えるべきか、というタグつき句分割推定問題を考える。$\boldsymbol{U}_{1:m}$ の生起確率は、その可能な全ての分割方法を考慮すると以下のようになる。

$$P(\boldsymbol{U}_{1:m}) = \sum_{\boldsymbol{V}, U(\boldsymbol{V})=\boldsymbol{U}_{1:m}} P(\boldsymbol{V}) = \sum_{k=1}^{m-1} P(\boldsymbol{U}_{1:k-1})\, P(v_{k:m}) \tag{1}$$

ただし、$v_{i:j}$ はタグつきテキスト $\boldsymbol{U}_{i:j}$ からなる 1 個のタグつき句、$U(\boldsymbol{V})$ はタグつき句列 $\boldsymbol{V}$ を結合したタグつきテキストを表す。従って、最尤なタグつき句列やタグつき句列のサンプルは、動的計画法を用いて高速に求めることができる [4]。

テキストへのタグづけの推定方法　テキスト（文字列）$C_{1:n} = c_1 \cdots c_n$ に対する最尤タグ列の推定は、表層が $C_{1:n}$ となる最尤のタグつきテキスト

$$\hat{\boldsymbol{U}}_{\boldsymbol{C}_{1:n}} = \operatorname*{arg\,max}_{\boldsymbol{U}, C(\boldsymbol{U})=\boldsymbol{C}_{1:n}} P(\boldsymbol{U}) \tag{2}$$

（$C(\boldsymbol{U})$ はタグつきテキスト $\boldsymbol{U}$ の表層を表す）を求めることに等しいが、表層が $C_{1:n}$ となる最尤タグつき句列 $\hat{\boldsymbol{V}}_{\boldsymbol{C}_{1:n}}$ の探索の方が高速であるので、これで代用することとする。$\hat{\boldsymbol{V}}_{\boldsymbol{C}_{1:n}}$ の生起確率は

$$\begin{aligned} P(\hat{\boldsymbol{V}}_{\boldsymbol{C}_{1:n}}) &= \max_{\boldsymbol{V}_{\boldsymbol{C}_{1:n}}} P(\boldsymbol{V}_{\boldsymbol{C}_{1:n}}) \\ &= \max_{1\le l\le n-1} \left(P(\hat{\boldsymbol{V}}_{\boldsymbol{C}_{1:l-1}}) \cdot \max_{v, C(v)=C_{l:n}} P(v) \right) \end{aligned} \tag{3}$$

となるので、$\hat{\boldsymbol{V}}_{\boldsymbol{C}_{1:n}}$ は動的計画法を用いて高速に求めることができ、タグづけ結果を得ることができる。

4 Dependent Dirichlet Process

あらかじめ学習データが固定的に与えられた場合、Dirichlet process (DP) を用いたタグつき句モデル学習は、Gibbs sampler などを用いてデータ全体の分割を反復推定することで実現できる。しかし我々は、学習データが 1 つ増えるたびに直ちに学習を行って、学習結果を直後のタグづけタスクに反映したいため、学習データ全体の反復推定による学習は適さない。また、1 章で述べたような「ある時点からつけられるタグが変更になる」ことを反映したモデルを構築することは、学習データ全体を一様に扱う枠組みでは難しい。

Araki らは「忘却」を用いた DP のオンライン学習手法を提案している [1]。これは、記憶している以前の推定結果を反復再推定する代わりに、一定時間の後に忘却することで最適化を図る手法であり、タグの変更への対応にも一定の効果が期待できるため、我々の目的に適している。しかし、変更への追従性を高めるためには忘却を早める必要があり、性能上問題がある。

そこで我々は、Lin らが提案している dependent Dirichlet process (DDP) [3] を用いることで、このような忘却を実現し、かつ、記憶の「書き換え」を行うモデルを構築することにした。DDP は DP の Markov 連鎖として構築され、時刻を進める際に記憶の確率的な忘却や書き換えを行うことができる。

以下では、Chinese restaurant process の用語を用いて本稿で提案する手法で用いる DDP モデルを説明する。前時刻 $t-1$ において、基底分布 G_{t-1}、集中パラメタ α_{t-1} の DP $D_{t-1} \sim \mathrm{DP}(\alpha_{t-1}G_{t-1})$ があり、また $t-1$ までのタグつき句の観測結果がレストラン R（テーブル数は m、テーブル t_k の客数は c_k、料理は v_k）に保持されており、加えて各テーブル t_k にはその存在確率 e_k が与えられているものとする。

$$D_{t-1}|R \sim \mathrm{DP}\left(\alpha_{t-1}G_{t-1} + \sum_{i=1}^{m} e_i c_i \delta_{v_i}\right) \tag{4}$$

ただし、δ_v は v に質点を置く Dirac の delta である。

ここで時刻を t に進め、基底分布 G_t および集中パラメタ α_t を持つ新たな DP $D_t \sim \mathrm{DP}(\alpha_t G_t)$ を考える。また、時刻が進んだ際に、R の内容に確率的に以下の変更が起こったと考える。

1. （確率 e_k で存在している）各テーブルは、各々保持確率 q の Bernoulli 分布 Q_k に従って保持、もしくは客ごと破棄される。

2. 1. で保持された各テーブルの料理は、置換確率 $T_t(v_k, v')$ に従って、従来のタグつき句 v_k から v' に置換される（$v' = v_k$ なら置換されない）。

この R に対する D_t の事後分布は以下の通り（ただし $T_t(v)$ は v の置換先についての確率分布であり、$P(T_t(v) = v') = T_t(v, v')$）。

$$D_t|R \sim \mathrm{DP}\left(\alpha_t G_t + \sum_{i=1}^{m} q e_i c_i T_t(v_i)\right) \quad (5)$$

この R を時刻 t の観測結果を反映して更新していく。タグつき句が生起する期待確率は

$$P(v \sim D_t|R) = \frac{\alpha_t G_t(v) + \sum_{i=1}^{m} q e_i c_i T_t(v_i, v)}{\alpha_t + \sum_{i=1}^{m} q e_i c_i} \quad (6)$$

となるので、観測したタグつき句が v であったならば、以下の更新結果 R' を得る（ただし $a = \alpha_t G_t(v) + \sum_{i=1}^{m} q e_i c_i T_t(v_i, v)$）。

- 確率 $\alpha_t G_t(v)/a$ で、料理が v の新規テーブル t_{m+1} を作成し、客を配する（$c_{m+1} = e_{m+1} = 1$）。

$$D_t|R' \sim \mathrm{DP}\left(\alpha_t G_t + \sum_{i=1}^{m} q e_i c_i T_t(v_i) + \delta_v\right) \quad (7)$$

- 確率 $q e_k c_k T_t(v_k, v)/a$ で、既存のテーブル t_k に客を追加する。このとき同時に、時刻 $t-1$ の時点で t_k が存在し（$e_k \leftarrow 1$）、かつ時刻 t にて引き続き保持され（$Q_k \leftarrow \{\text{保持} = 1\}$）、さらに料理が v に置換された（$T_t(v_k) \leftarrow \delta_v$）ことが確定する。

$$D_t|R' \sim \mathrm{DP}\left(\alpha_t G_t + \sum_{i=1,\ldots,m \wedge i \neq k} q e_i c_i T_t(v_i) + (c_k + 1)\delta_v\right) \quad (8)$$

同時刻の以降の観測も同様の手順で行い、全ての観測結果の反映が終了した後で、この時刻に客が追加されなかった（テーブルの保持や料理の置換先が未確定の）各テーブル t_k について、$e_k \leftarrow q e_k$ とし、またタグつき句の置換先を $T_t(v_t)$ から sampling する。

5 提案手法

前章の DDP を用い、学習対象のタグつき文書が 1 つ得られる度に時刻を進めてこの文書に対する観測を行う（時刻 t にはただ 1 つの文書 d_t を学習する）ことで、タグつき句生起モデルのオンライン学習を行う。

各種分布や確率の与え方　時刻 t の基底分布 G_t は、$\{d_1, \ldots, d_t\}$ から得られるタグ unigram 多項分布モデル P_{ut} を用い、以下のように与えることとする[4]（$|v|$ は v 中のタグ数。s はタグつき句長パラメタで固定的に与える）。

$$G_t(v) = (1-s)^{|v|-1} s \prod_{u \in v} P_{\mathrm{ut}}(u) \quad (9)$$

[4] タグつき句モデルの学習に先立ってこのデータを含んだタグモデルを作成するので、タグつき句モデル学習時には常に $G_t(v) > 0$。

また集中パラメタ α_t は、D_t を新規に作成した時点での R の内容から推定する [6]。

保持確率 q は学習を通じて固定値を与える。一方、タグつき句 v の v' への置換確率 $T_t(v, v')$ は d_t に依存して各時刻で異なる値を与えることとする。具体的には、$v = \langle w, T\rangle$ に対して、d_t 中に同じ表層 w に対して異なるタグ列 T' が付与されたものが含まれているならば、一定の確率 r で v が $v' = \langle w, T'\rangle$ に置換される（$T_t(v, v') = r, T_t(v, v) = 1-r, T_t(v, \neg v \wedge \neg v') = 0$）こととし、さもなくば置換は起こらない（$T_t(v, v) = 1$, $T_t(v, \neg v) = 0$）こととする。

形態素解析との併用　本手法では、「表層」は文字列であって問題ないが、我々は形態素解析器によって解析された形態素列を用いることとした。解析結果の各形態素には品詞等の曖昧性解消に有用な情報が付与されているので、タグづけ性能の向上が期待できる。また後述のように、タグ推定時に未知語に対して、解析結果の品詞や読み等の情報に基づいてタグを自動生成することが可能となる。

学習手順

0. 時刻 $t = 0$ の空の R を用意する。
1. 新たなタグつき文書を入手したら $t \leftarrow t+1$ として、以下の手順でこの文書 d_t からの学習を実施:
   (a) R を参照して α_t を推定
   (b) d_t を形態素解析し、各タグと形態素列とを対応づけ
   (c) $d_1, \ldots d_t$ に含まれるタグの出現頻度を集計し、タグ生起モデル P_{Ut} を計算
   (d) d_t より T_t を決定
   (e) d_t のタグつき句列への分割を推定（3 章）（文書全体を 1 block とした blocked sampling を行うこととする。）
   (f) (e) の結果観測された各タグつき句を R に反映。また客が追加されなかった各テーブルの存在確率とタグつき句を更新（4 章）。
2. この時点の R は最新の学習結果として、タグづけ推定に供することができる。1. に戻る。

タグづけ　タグづけの際には、最新の R より求められる、基底測度成分を除いたタグつき句生起確率 $P'(v|R) = \sum_{k, l_k = v} e_k c_k / \sum_i e_i c_i$（$l_k$ は t_k のタグつき句）を用いることとする。入力文を形態素解析した後、3 章で述べたように動的計画法を用いて形態素列を覆う最大確率のタグつき句列を探索し、得られたタグ列を出力する。ただし、R に入力文を覆うことのできるタグつき句が存在しない可能性があるため、任意の形態素 1 個を含むタグ 1 個から構成されるタグつき句（デフォルトタグつき句）が極小確率で生起するものとして探索を行うこととする。このタグの読みや種別は、形態素に付与されている情報から自動的に生成することができる。また、あらかじめ「形態素→タグ対応表」を用意しておき、併用することも可能である。

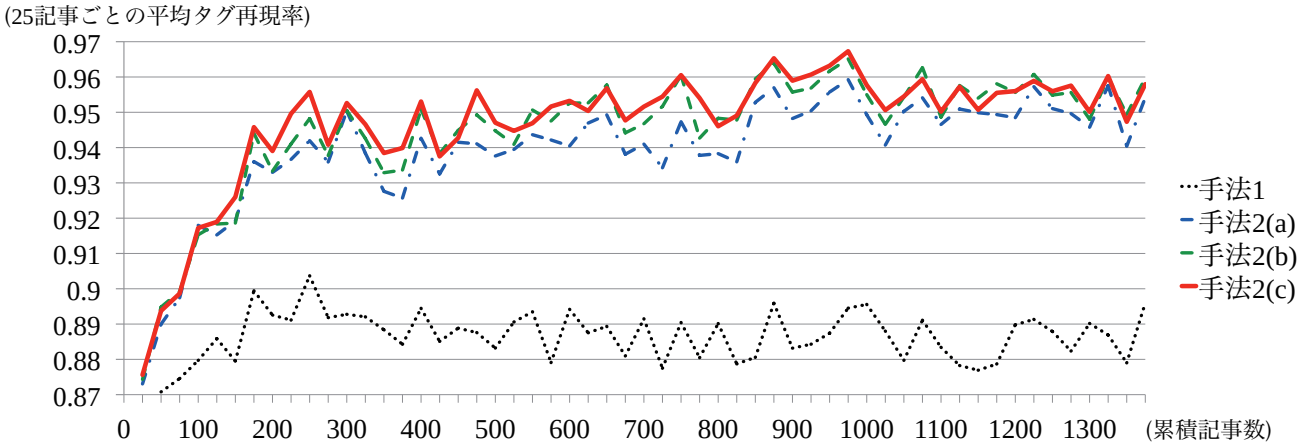

図 1: 学習の経過に伴う各手法のタグ再現率の推移

6 実験・考察

NWE サービスで作成し人手によるタグつきやさしい日本語ニュース原稿 1,372 記事を用いて、タグづけ性能評価実験を行った。作成日時の古い記事から順に、その時点のモデルを用いた記事への自動タグづけ結果を人手の結果と比較し、その後この記事の学習を行う、という操作を繰り返した。

以下に挙げる手法の間で、正解データ（人手のタグ）のタグ再現率を尺度に性能比較を行った。

1. 比較手法
   現状我々が NWE 作成現場に提供している、形態素解析（MeCab + IPAdic）と形態素→タグ対応表を用いる手法。
2. 提案手法
   (a) タグつき句長を 1 に固定、タグ書き換えあり
   (b) タグつき句長は任意、タグ書き換え なし
   (c) タグつき句長は任意、タグ書き換えあり
   形態素解析器 MeCab + UniDic を併用し、また手法 1 で用いている形態素→タグ対応表[5] を用いてデフォルトタグつき句にタグを付与した。また、学習時パラメタは、保持確率 $q = 0.99$、タグ書き換え確率（(a), (c) のみ）$r = .01$、タグつき句長パラメタ（(b), (c) のみ）$s = 0.9$ を与えた。

図 1 に評価結果を示す。手法 1 は学習を行わないので性能はほぼ一定であるが、手法 2 はオンライン学習の進行につれて性能が向上した。また、手法 2(a) は 2(b)(c) に比べて性能が劣ることから、タグつき句を記憶単位とすることで文脈に応じたタグを付与する能力が向上しているといえる。

手法 2(b),(c) の性能差は微小だが、2(b) には学習の過程で性能が大きく低下している箇所があるのに対して、2(c) では同じ時期の性能低下がより小さいことがわかる。この時期のタグづけ結果を 2(b),(c) 間で比較してみた結果、頻出表現へのタグづけ方針の変更があったことがわかった。以上の分析結果から、タグ書き換え機能はタグの変更への追従に有効だと考える。

提案手法による誤りの多くは、学習データに含まれない、いわゆる未知語を含む部分であった。今後さらに性能を向上させるためには、未知語に対するタグの推定機能を導入する必要があるといえる。

[5] 形態素体系の違いは自動変換した。

7 おわりに

本稿では、dependent Dirichlet process を用いることで、日本語文書へのタグづけ知識をオンライン学習する手法を提案し、本手法を用いたタグづけ能力が日々タグづけ作業を行う現場の支援に有効であることを示した。今後は、特に未知語に対するタグづけ性能を向上させるため、多くの形態素解析器が行っているような各種未知語処理の導入を検討することを考えている。また、タグつき句 unigram をより高次の n-gram に拡張することも検討したい。

参考文献

[1] T. Araki, T. Nakamura, T. Nagai, K. Funakoshi, M. Nakano, and N. Iwahashi. Autonomous acquisition of multimodal information for online object concept formation by robots. In *Proceedings of IROS 2011*, pp. 1540–1547, 2011.

[2] 国際交流基金, 日本国際教育支援協会（編）. 日本語能力試験出題基準（改訂版）. 凡人社, 1994.

[3] D. Lin, E. Grimson, and J. Fisher. Construction of dependent Dirichlet processes based on Poisson processes. In *Proceedings of NIPS 2010*, 2010.

[4] 持橋, 山田, 上田. ベイズ階層言語モデルによる教師なし形態素解析. 情報処理学会研究報告 2009-NL-190, 2009.

[5] 田中, 美野, 越智, 柴田. やさしい日本語ニュースの公開実験. NHK 技研 R&D, No. 139, pp. 20–29, 2013.

[6] M. West. Hyperparameter estimation in Dirichlet process mixture models. Technical report, Institute of Statistics and Decision Sciences, Duke University, 1992.