Vol.2013-NL-213 No.10
2013/9/13

ベイズ階層言語モデルとSemi-Markov SHDCRFの協調学習による教師なし形態素解析

内海 慶[1,a)]　塚原 裕史[1,b)]

概要：本論文では，教師なし・半教師あり学習による形態素解析手法の提案を行う．従来の教師なし形態素解析手法では分かち書きのみを対象としており，品詞推定は扱っていなかった．我々は，この問題に対処するため，潜在クラスを導入したSemi-Markov CRFとNPYLMの協調学習を行った．新聞データ及びブログ記事を用いた実験によって，提案手法の有効性を評価した．

キーワード：条件付き確率場，言語モデル，形態素解析，ノンパラメトリックベイズ

1. はじめに

形態素解析は自然言語処理における重要な基盤技術である．特に，日本語や中国語のような分かち書きされない言語においては，文書分類や文書検索の索引付け，固有表現抽出といった様々な自然言語処理タスクの前処理として使用される．JumanやChaSen，MeCab等，従来言語処理で利用されてきた形態素解析は書き言葉を対象として作られてきた．しかし，近年ではブログやSNS，Twitter等のCGMが増加しており，これらの一般消費者が生成するメディアを対象とした評判情報等の研究も行われている．

CGMでは書き言葉と話し言葉が混同されて用いられる．例えば，文頭の接頭語，「なのに」という表現は「それなのに」の省略形であり，本来は書き言葉ではない．しかしながら，例えばTwitterで「なのに」と検索をすると接頭語として用いているツイートを見ることができる．また，書き言葉では見られない顔文字を用いた感情の表現や「△」等の記号の読みと口語を対応させた表現等も見られる．

このように，話し言葉を含む文章では，文法の誤りや新聞記事には現れない未知語が現れる．また，話し言葉は変化が早く，例えば90年代半ばに流行したギャル語は，現在では死語となった物も多く，また一方で新たな表現も産まれている．

従来の形態素解析のように，話し言葉について正解データを作るのは難しく，仮に人手で多量の正解データを作れたとしても常に現れ続ける新語に対応し続けるのは現実的ではない．

こうした問題に対して，教師データなしや，既存の教師データを利用しつつ，話し言葉の分かち書きを行う教師なし学習や半教師あり学習の手法が提案されている．しかしながら，これまで提案されてきた手法では，分かち書きは獲得できるものの，品詞情報の獲得は対象とされていなかった．品詞情報は固有表現抽出や係り受け解析等，形態素解析を前処理として用いる解析では重要な手がかりとして利用される．

そこで，本論文では，教師なし，半教師あり学習で話し言葉の形態素解析の学習を行うと同時に，品詞情報の獲得を行う手法を提案する．

以降，2章では話し言葉を対象とした形態素解析に関連する選考研究について説明を行い，3章で我々が基にした持橋らのNPYCRF[1]とShenらのSHDCRF[2]について解説する．4章で，我々の提案する形態素解析手法について述べる．5章では，我々の手法について評価を行い，その効果を示す．6章では，総論を行い，今後の課題を示す．

2. 関連研究

話し言葉の形態素解析においては，未知語が解析の問題となることが指摘されている[3]．内元らは少量のタグ付きコーパスから，最大エントロピーモデルを用いて文字列の単語と品詞の同時推定を行い，尤度の低い形態素を人手で修正することで効率よく未知語に対する情報を付与している．松本ら[4]は，既存の書き言葉を対象に作られた形態素解析が話し言葉では性能が出ないことを示し，これに

[1] デンソーアイティーラボラトリ
DENSO IT LABORATORY, cross tower 28th Floor, 2-15-1 Shibuya Shibuya-ku Tokyo, 150–0002, Japan
[a)] kuchiumi@d-itlab.co.jp
[b)] htsukahara@d-itlab.co.jp

対して少量のタグ付き話し言葉データを既存の書き言葉の教師データに加えることで，大きく性能を改善できることを示した．これらの手法は教師あり学習に基づく手法であり，形態素解析の学習をするためには教師データを必要とする．しかし，前述したように話し言葉は変化が早く，常に変化する表現にあわせて教師データを作成し続けるのは現実的ではない．

Creutz ら [5] は，英語及びフィンランド語の単語分割に，最小記述長 (MDL) に基づくグリーディアルゴリズムを用いる手法と，EM アルゴリズムで用いて入力データの対数尤度を最大化する手法の 2 つを提案し，MDL に基づく手法がより高い正解率となることを示した．Argamon ら [6] も同様に，MDL に基づくグリーディアルゴリズムによって単語分割を行っている．Argamon らは，単語分割前と後での MDL の差分の式を導出することで計算量の削減を行った．松原ら [7] は，Argamon らの手法をベースに，単語分割では無く，文字のチャンキングを行うことで日本語話し言葉の単語分割を行った．MDL に基づくチャンキングでは，チャンキングを繰り返すうちにデータに対して過学習が起こることが問題として挙げられている．また，計算コストが非常に大きく，少量のコーパスしか扱えない．

持橋ら [8] は，文字・単語ベイズ n グラム言語モデルのベイズ学習を用い，言語に依存しない単語分割の提案を行った．彼らの手法では，最適なスムージングを行うベイズ n グラム言語モデルを用いることで，学習データに対する言語モデルの過学習の問題を解決している．しかし，言語モデルの性能を最適化しているため，人間の分割基準とは異なる [*1] 場合があった．この問題に対処するため，持橋らは，CRF と NPYLM の協調学習を行う半教師有り学習手法による単語分割も提案している [1].

これまで提案された教師あり学習に基づく形態素解析手法では，話し言葉を扱う際にも教師データを必要としており，また教師なし学習による手法では分かち書きは扱っているものの，単語の潜在的な意味クラスを考慮していないため品詞推定は行えなかった．

我々の提案する手法では，単語の潜在的な意味クラスを考慮し，単語分割と同時にその推定を行う．これまでの手法では単語分割のみが対象であったが，提案手法では単語の意味クラスと，意味クラスの間の依存関係をデータから獲得することで，単語の意味，すなわち品詞と，品詞間の依存関係，すなわち文法を獲得する．

3. 従来手法

我々の提案する手法は持橋らの半教師あり形態素解析手法を基にしている．ここでは最初に，持橋らの NPYCRF について説明する．

[*1] 「見/る」のように活用語尾が分割されたり，複合語が 1 単語として切り出されてしまう．

3.1 NPYCRF

持橋らの手法では，鈴木らの JESS-CM 法 [9] を用いて，単語分割の生成モデルである NPYLM を CRF による形態素解析の識別モデルと統合している．JESS-CM では，入力 $\mathbf{x}$ に対するラベル $\mathbf{y}$ の条件付き確率を以下で表現する．

$$p(\mathbf{y}|\mathbf{x}) \propto p_{DISC}(\mathbf{y}|\mathbf{x};\mathbf{\Lambda})p_{GEN}(\mathbf{y},\mathbf{x};\mathbf{\Theta})^{\lambda_0} \tag{1}$$

p_{DISC} は識別モデル，p_{GEN} は生成モデルであり，$\mathbf{\Lambda}$，$\mathbf{\Theta}$ はそれぞれ識別モデル，生成モデルのパラメータである．(1) 式の識別モデルを CRF のような対数線形モデル

$$p_{DISC}(\mathbf{y}|\mathbf{x};\mathbf{\Lambda}) \propto \exp(\sum_{k=1}^{K} \lambda_k f_k(\mathbf{y},\mathbf{x})) \tag{2}$$

にとれば，(1) 式は $\log p_{GEN}(\mathbf{y},\mathbf{x};\mathbf{\Theta})$ を 1 つの素性関数とみて，

$$p(\mathbf{y}|\mathbf{x};\mathbf{\Theta}) \propto \exp(\lambda_0 \log p_{GEN}(\mathbf{x},\mathbf{y}) + \sum_{k=1}^{K} \lambda_k f_k(\mathbf{y},\mathbf{x})) \tag{3}$$

$$= \exp(\mathbf{\Lambda} \cdot \mathbf{F}(\mathbf{y},\mathbf{x})) \tag{4}$$

と，パラメータ $\mathbf{\Lambda} = (\lambda_0, \lambda_1, \ldots, \lambda_K)$ を持つ対数線形モデルの形でも書くことができる．(1) 式と (3) 式は等価であるから，JESS-CM ではこの 2 つの式を用いて，ラベル付きデータ $(\mathbf{X_l}, \mathbf{Y_l})$ 及びラベルなしデータ $(\mathbf{X_u}, \mathbf{Y_u})$ が与えられた時に，以下の目的関数

$$p(\mathbf{X_u}, \mathbf{Y_l}|\mathbf{X_l};\mathbf{\Lambda},\mathbf{\Theta}) = p(\mathbf{Y_l}|\mathbf{X_l}) \cdot p(\mathbf{X_u}) \tag{5}$$

の値を，$\mathbf{\Lambda}$ 及び $\mathbf{\Theta}$ について交互に最大化する協調学習を行う．

JESS-CM では CRF-HMM の半教師学習を行っており，2 つの学習器は同じ構造を持つことを前提としているが，持橋らの NPYCRF では，NPYLM が Semi-Markov モデルの構造を持つのに対して，CRF は Markov モデルの構造となっている．そのため，Markov モデルと Semi-Markov モデルの情報を相互に変換することで，違う構造を持つモデル間の半教師あり学習を行っている．

我々の提案する手法でも，持橋らと同様に JESS-CM の枠組を用いて NPYLM と CRF の協調学習を行う．しかし，我々の手法では品詞推定を同時に行うために，CRF に潜在変数を導入した手法を用いる．また，持橋らと異なり，Semi-Markov モデルから Markov モデルへの変換は行わない．以降で，潜在変数を導入した CRF について説明する．

3.2 SHDCRF

品詞推定と分かち書きを同時に行うために，我々は Shen らの SHDCRF[2] を用いる．SHDCRF は検索クエリのユー

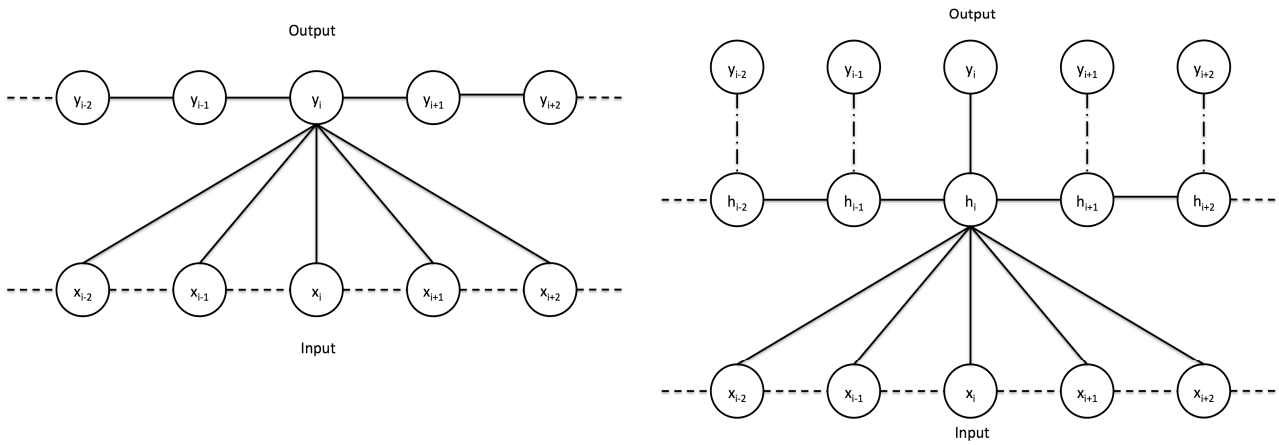

図 1　CRF と SHDCRF のグラフ構造
Fig. 1　Structure of CRF and SHDCRF

ザ意図を推定するために提案された手法で，図 1 に示すように入力系列 $\mathbf{x}$ とラベル列 $\mathbf{y}$ との間に潜在変数 $\mathbf{h}$ が導入されており，出力ラベル $\mathbf{y}$ の依存関係ではなく，潜在変数 $\mathbf{h}$ の間の依存関係を考慮したモデルとなっている．SHDCRF では，条件付き確率 $p(\mathbf{y}|\mathbf{x})$ は以下のようにモデル化される．

$$p_\Lambda(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{h}} p_\Lambda(\mathbf{y}|\mathbf{h}) p_\Lambda(\mathbf{h}|\mathbf{x}) \tag{6}$$

$$p_\Lambda(\mathbf{y}|\mathbf{h}) = \frac{1}{z_1(\mathbf{h})} \exp(\sum_{k=1}^{p} \beta_k G_k(\mathbf{y}, \mathbf{h})) \tag{7}$$

$$p_\Lambda(\mathbf{h}|\mathbf{x}) = \frac{1}{z_2(\mathbf{x})} \exp(\sum_{k=1}^{q} \lambda_k F_k(\mathbf{h}, \mathbf{x})) \tag{8}$$

ここで，$z_1(\mathbf{h})$ 及び $z_2(\mathbf{h})$ は規格化のための分配関数を，G_k 及び F_k は潜在クラスの系列とラベル列，入力系列に対する素性関数をそれぞれ表す．λ_k と β_k は素性関数に対応するモデルパラメータであり，$\Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_p, \beta_1, \beta_2, \ldots, \beta_q)$ である．G_k，F_k は系列に対する素性関数であり，以下で表される．

$$G_k(\mathbf{y}, \mathbf{h}) = \sum_{i=1}^{T} g_k(y_i, h_i) \tag{9}$$

$$F_k(\mathbf{h}, \mathbf{x}) = \sum_{i=1}^{T} f_k(h_{i-1}, h_i, \mathbf{x}) \tag{10}$$

$g_k(y_i, h_i)$ は各位置での出力ラベルと潜在クラスに関する素性関数，$f_k(h_{i-1}, h_i, \mathbf{x})$ は，各位置での観測素性 $\mathbf{x}$ に関する素性関数 $s_k(h_i, \mathbf{x})$ 及び，潜在クラスの遷移素性に関する素性関数 $t_k(h_{i-1}, h_i, \mathbf{x})$ を含む．SHDCRF では，潜在クラスについての遷移素性関数 $t_k(h_{i-1}, h_i, \mathbf{x})$ によって潜在クラスの変化を学習する．

3.2.1　SHDCRF のパラメータ推定

SHDCRF では，以下の損失関数を最大化することでパラメータの学習を行う．

$$L(\Lambda) = \sum_{(\mathbf{x},\mathbf{y})} \tilde{p}(\mathbf{x}, \mathbf{y}) \log p_\Lambda(\mathbf{y}|\mathbf{x}) - \frac{||\Lambda||^2}{2\sigma^2} - \alpha H_\Lambda(Y|H) \tag{11}$$

$$H_\Lambda(Y|H) = -\sum_{\dot{h} \in H} \sum_{\dot{y} \in Y} p_\Lambda(\dot{y}|\dot{h}) \log p_\Lambda(\dot{y}|\dot{h}) \tag{12}$$

最初の 2 項は CRF の損失関数と同様である．第 3 項はラベルの潜在クラスに対する条件付きエントロピーである．SHDCRF では，この条件付きエントロピーを最小化することで，潜在クラスが与えられた際のラベルの曖昧さを減らしている．これによって，ラベルに紐づく潜在クラスが疎になり，ラベルと特徴的な潜在クラスのみがサブクラスとして割り当てられる．

4.　提案手法

ここでは我々の提案手法について述べる．3.2 で説明した SHDCRF は Markov モデルであり，Semi-Markov モデルである NPYLM とは構造が異なるため，JESS-CM の枠組みにそのままでは適用できない．そのため，我々の提案手法では Sarawagi ら [10] の Semi-Markov CRF モデルを基に，Semi-Markov モデルの SHDCRF を用いる．セグメントに対して 1 つの潜在クラスが割り当てられ，それは単語の意味クラスと考えることができるため，ここではそれを品詞と見なす．

単純に Semi-Markov モデルを適用し，セグメント素性に NPYLM を利用しただけでは，品詞に関する素性関数がどの品詞クラスに対しても等しい戻り値となり，そのままでは分かち書きはできても品詞推定はうまくいかない．また，セグメント素性にセグメントに含まれる文字列を用いた場合には，品詞推定は行えるが，IOB タグや，Begin, Continue 等のラベルを用いた Markov モデルでは捉えられる，セグメントの先頭になりやすい文字列や文末になりやすい文字列，またはセグメントの先頭になりにくい文字列

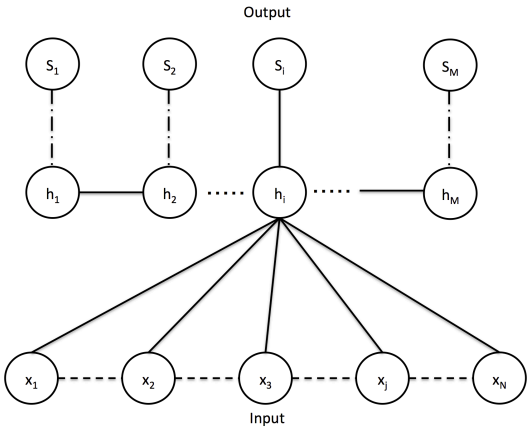

図 2 Semi-Markov SHDCRF のグラフ構造 (M ≤ N)
Fig. 2 Structure of Semi-Markov SHDCRF(M ≤ N)

や文末になりにくい文字列といった特徴は，Semi-Markov モデルでは捉えられない [*2]．分かち書きを行う上で，IOB タグのようにセグメントの開始や終了位置になりやすい，またはなりにくいという特徴を扱えることは非常に強力なアドバンテージだと考えられる．

そこで，我々は Galen ら [11] の提案する Hybrid Markov/Semi-Markov CRF のアイデアを取り込むことで Semi-Markov SHDCRF で Markov 素性を扱い，文字列の開始位置や終了位置へのなりやすさ，品詞との関係を捉えられるように拡張した．

4.1 Semi-Markov SHDCRF

Semi-Markov SHDCRF では，入力文字列 $\mathbf{x}$ が与えられた際の形態素解析結果 $\mathbf{s}$ が得られる確率を以下で表す．

$$p_\Lambda(\mathbf{s}|\mathbf{x}) = \sum_{\mathbf{h}:|\mathbf{h}|=|\mathbf{s}|} p_\Lambda(\mathbf{s}|\mathbf{h}) p_\Lambda(\mathbf{h}|\mathbf{x}) \tag{13}$$

ただし，$\mathbf{h} = \{h_1, h_2, \ldots, h_M\}$ の時，$|\mathbf{h}| = M$ とする．

Semi-Markov SHDCRF のグラフ構造を，図 2 に示す．s_i は，潜在クラスの列 $\mathbf{h}$ に対応するセグメンテーションに含まれるセグメントであり，h_i と 1 対 1 で対応する．

$p_{\Lambda(\mathbf{s}|\mathbf{x})}$ の各因子 $p_\Lambda(\mathbf{s}|\mathbf{h})$，$p_\Lambda(\mathbf{h}|\mathbf{x})$ を以下に表す．

$$p_\Lambda(\mathbf{s}|\mathbf{h}) = \frac{1}{z_1(\mathbf{h})} \exp(\sum_k \beta_k \Psi_k(\mathbf{s},\mathbf{h}) + \Psi(\mathbf{s},\mathbf{h})) \tag{14}$$

$$p_\Lambda(\mathbf{h}|\mathbf{x}) = \frac{1}{z_2(\mathbf{x})} \exp(\sum_k \lambda_k \Phi_k(\mathbf{x},\mathbf{h}) + \Phi(\mathbf{x},\mathbf{h})) \tag{15}$$

ただし，

$$\Psi(\mathbf{s},\mathbf{h}) = \begin{cases} 0 & |\mathbf{s}| = |\mathbf{h}| \\ -\infty & |\mathbf{s}| \neq |\mathbf{h}| \end{cases} \tag{16}$$

$$\Phi(\mathbf{h},\mathbf{x}) = \begin{cases} 0 & |\mathbf{h}| \leq |\mathbf{x}| \\ -\infty & |\mathbf{h}| > |\mathbf{x}| \end{cases} \tag{17}$$

である．

$z_1(\mathbf{h})$ 及び $z_2(\mathbf{x})$ は規格化のための分配関数で，以下の式で表される．

$$\begin{aligned} z_1(\mathbf{h}) &= \sum_{\mathbf{s}} \exp(\sum_k \beta_k \Psi_k(\mathbf{s},\mathbf{h}) + \Psi(\mathbf{s},\mathbf{h})) \\ &= \sum_{\mathbf{s}:|\mathbf{s}|=|\mathbf{h}|} \exp(\sum_k (\beta_k \Psi_k(\mathbf{s},\mathbf{h}))) \end{aligned} \tag{18}$$

$$\begin{aligned} z_2(\mathbf{x}) &= \sum_{\mathbf{h}} \exp(\sum_k \lambda_k \Phi_k(\mathbf{x},\mathbf{h}) + \Phi(\mathbf{x},\mathbf{h})) \\ &= \sum_{\mathbf{h}:|\mathbf{h}|\leq|\mathbf{x}|} \exp(\sum_k \lambda_k \Phi_k(\mathbf{x},\mathbf{h})) \\ &= \sum_{M=1}^{|\mathbf{x}|} \sum_{\mathbf{h}:|\mathbf{h}|=M} \exp(\sum_k \lambda_k \Phi_k(\mathbf{x},\mathbf{h})) \end{aligned} \tag{19}$$

$\Phi_k(\mathbf{x},\mathbf{h})$ はセグメンテーションに対応した潜在変数の系列と入力系列に関する素性関数となっている．

素性関数は，より詳しくは

$(k \neq 0)$

$$\Psi_k(\mathbf{s},\mathbf{h}) = \sum_{j=1}^{|\mathbf{s}|} \psi_k(s_j, h_j) \tag{20}$$

$$\Phi_k(\mathbf{x},\mathbf{h}) = \sum_{j=1}^{|\mathbf{h}|} \phi_k(h_{j-1}, h_j, \mathbf{x}) \tag{21}$$

$(k = 0)$

$$\Phi_0(\mathbf{x},\mathbf{h}) = \log p_{GEN}(\mathbf{s}(\mathbf{h}), \mathbf{x}) \tag{22}$$

である．$p_{GEN}(\mathbf{x}, \mathbf{s}(\mathbf{h}))$ は NPYLM，$\mathbf{s}(\mathbf{h})$ は $\mathbf{h}$ に対応するセグメンテーションを表す．我々はこれを，さらに以下のように Markov 素性関数の和の形で再定義する．

$$\phi_k(h_{j-1}, h_j, \mathbf{x}) = \sum_{i=s(s)}^{e(s)} \phi_k^M(h_{i-1}, h_i, \mathbf{x}) \tag{23}$$

j はセグメンテーションが与えられた際のセグメントの位置を，i は Markov モデルで見た際の入力系列の各位置を表す．Semi-Markov SHDCRF では各セグメントに対して 1 つの潜在クラスが割り当てられており，Markov モデルのラベルのようにローカルのラベルに分解できないが，素性関数を (24) 式のように開始位置，セグメントの内部，終了位置に関するものの 3 つに分解することで IOB タグを用いた際と同様の情報を扱えるようになる．s はセグメントの開始位置，i はセグメント中の位置 $s < i < e$ を，e は

[*2] SHDCRF の場合，潜在クラスと素性関数の間の関係のみを見ているため，Markov モデルを用いても単純には捉えられない．

セグメントの終了位置をそれぞれ表す．

$$\phi_k^M(h_{j-1},h_j,\mathbf{x}) = \begin{cases} \phi_{BEGIN}^M(s,h_{j-1},h_j,\mathbf{x}) & \\ \quad + \phi_{END}^M(s,h_{j-1},h_j,\mathbf{x}) & (segment\ size=1) \\ \phi_{BEGIN}^M(s,h_{j-1},h_j,\mathbf{x}) & \\ \quad + \phi_{END}^M(e,h_{j-1},h_j,\mathbf{x}) & (segment\ size=2) \\ \phi_{BEGIN}^M(s,h_{j-1},h_j,\mathbf{x}) & \\ \quad + \sum_{i=s+1}^{e-1}\phi_{INSIDE}^M(i,h_{j-1},h_j,\mathbf{x}) & \\ \quad + \phi_{END}^M(e,h_{j-1},h_j,\mathbf{x}) & (segment\ size>2) \end{cases} \tag{24}$$

4.2 パラメータ推定

Semi-Markov SHDCRF の損失関数を以下に表す．

$$L(\Lambda) = \sum_{(\mathbf{x},\mathbf{s})} p(\mathbf{x},\mathbf{s})\log p_\Lambda(\mathbf{s}|\mathbf{x}) - \frac{||\Lambda||^2}{2\sigma^2} - \alpha H_\Lambda(S|H) \tag{25}$$

$\mathbf{s}$ は出力となるセグメンテーションを，$H_\Lambda(S|H)$ は以下に示す条件付きエントロピーを表す．

$$H_\Lambda(S|H) = -\sum_{\mathbf{h},\mathbf{s}} p_\Lambda(\mathbf{s}|\mathbf{h})\log p_\Lambda(\mathbf{s}|\mathbf{h}) \tag{26}$$

$\tilde{p}(\mathbf{x},\mathbf{s})$ は，$(\mathbf{x},\mathbf{s})$ の組についての経験的な確率分布 $\tilde{p}(\mathbf{x},\mathbf{s}) \approx 1/N\sum_{i=1}^n \delta_{\mathbf{x}\mathbf{x}^i}\delta_{\mathbf{s}\mathbf{s}^i}$ である．損失関数 L を最大化することで，パラメータ Λ の推定を行う．損失関数 L は隠れ層を含んでおり，凸関数ではない．そのため，得られる解は局所最適解であることに注意されたい．

損失関数の最大化には，勾配法を用いる．以下に，パラメータ λ_k，β_k についての勾配を示す．

$$\frac{\partial L(\Lambda)}{\partial\lambda_k} = \sum_{\mathbf{x},\mathbf{s}}\tilde{p}(\mathbf{x},\mathbf{s})\sum_{\mathbf{h}:|\mathbf{h}|=|\mathbf{s}|} p_\Lambda(\mathbf{h}|\mathbf{x},\mathbf{s})\Phi_k(\mathbf{x},\mathbf{h}) - \sum_{\mathbf{x}}\tilde{p}(\mathbf{x})\sum_{\mathbf{h}':|\mathbf{h}'|\le|\mathbf{x}|} p_\Lambda(\mathbf{h}'|\mathbf{x})\Phi_k(\mathbf{x},\mathbf{h}') - \frac{\lambda_k}{\sigma^2} \tag{27}$$

$$\frac{\partial L(\Lambda)}{\partial\beta_k} = \sum_{\mathbf{x},\mathbf{s}}\tilde{p}(\mathbf{x},\mathbf{s})\Bigg(\sum_{\mathbf{h}:|\mathbf{h}|=|\mathbf{s}|} p_\Lambda(\mathbf{h}|\mathbf{x},\mathbf{s})\Psi_k(\mathbf{s},\mathbf{h}) - \sum_{\substack{\mathbf{h}:|\mathbf{h}|=|\mathbf{s}|\\ \mathbf{s}':|\mathbf{s}'|=|\mathbf{h}|}} p_\Lambda(\mathbf{s}',\mathbf{h}|\mathbf{x},\mathbf{s})\Psi_k(\mathbf{s}',\mathbf{h})\Bigg) - \frac{\beta_k}{\sigma^2} + \alpha\sum_{\mathbf{s}}\sum_{\mathbf{h}:|\mathbf{h}|=|\mathbf{s}|} p_\Lambda(\mathbf{s}|\mathbf{h})\sum_{k'}\beta_{k'}\Psi_{k'}(\mathbf{s},\mathbf{h})\Bigg(\Psi_k(\mathbf{s},\mathbf{h}) - \sum_{\mathbf{s}':|\mathbf{s}'|=|\mathbf{h}|} p_\Lambda(\mathbf{s}'|\mathbf{h})\Psi_k(\mathbf{s}',\mathbf{h})\Bigg) \tag{28}$$

上記の勾配は，$\mathbf{x}$，$\mathbf{s}$，$\mathbf{h}$ がそれぞれ系列となっていて，そのままでは計算が困難である．そのため，以下に示すように系列を分解し，Forward-Backward アルゴリズムによって効率的に計算できるようにする．

$$\frac{\partial L(\Lambda)}{\partial\lambda_k} = \sum_{\mathbf{x},\mathbf{s}}\tilde{p}(\mathbf{x},\mathbf{s})\sum_{j=1}^{|\mathbf{s}|}\sum_{h_{j-1},h_j} p_\Lambda(h_{j-1},h_j|\mathbf{x}^i,\mathbf{s})\phi_k(h_{j-1},h_j,\mathbf{x}) - \sum_{\mathbf{x}}\tilde{p}(\mathbf{x})\sum_{j=1}^{|\mathbf{x}|}\sum_{h'_{j-1},h'_j} p_\Lambda(h'_{j-1},h'_j|\mathbf{x})\phi(h'_{j-1},h'_j,\mathbf{x}) - \frac{\lambda_k}{\sigma^2} \tag{29}$$

$$\frac{\partial L(\Lambda)}{\partial\beta_k} = \sum_{\mathbf{x},\mathbf{s}}\tilde{p}(\mathbf{x},\mathbf{s})\Bigg\{\sum_{j=1}^{|\mathbf{s}|}\Bigg(\sum_{h_j} p_\Lambda(h_j|\mathbf{x},\mathbf{s})\psi_k(h_j,s_j) - \sum_{h_j,s'_j} p_\Lambda(h_j,s'_j|\mathbf{x},\mathbf{s})\psi_k(h_j,s'_j)\Bigg)\Bigg\} - \frac{\beta_k}{\sigma^2} + \alpha\sum_{\mathbf{s}}\sum_{j=1}^{|\mathbf{s}|}\sum_{h_j} p_\Lambda(s_j|h_j)\sum_{k'}\beta_{k'}\psi_{k'}(s_j,h_j)\Bigg\{\sum_{j=1}^{|\mathbf{s}|}\Bigg(\psi_k(s_j,h_j) - \sum_{s'_j} p_\Lambda(s'_j|h_j)\psi_k(s'_j,h_j)\Bigg)\Bigg\} \tag{30}$$

Algorithm 1 Learning algorithm

Input: $x \in X$
Output: $\Lambda* = argmax_\Lambda L(\Lambda)$

```
Add X to NPYLM
t ← 1
while t ≤ LOOP do
    x ∼ X
    if t > 1 then
        Remove s*_{t−1} of x in NPYLM
    end if
    s ← argmax_s p_{Λ_t}(s|x)
    s*_t ∼ p_{Λ_t}(s|x)
    Λ_{t+1} ← Λ_t + η ∂Λ_t/∂λ_k |_s
    Λ_{t+1} ← Λ_t + η ∂Λ_t/∂β_k |_s
    Add s*_t to NPYLM
    t ← t + 1
end while
```

学習アルゴリズムには，NPYLM との協調学習の相性からオンライン学習を用いた．Algorithm 1 に，我々の学習アルゴリズムを示す．NPYLM の更新に用いるセグメンテーションは，持橋らと同様に現在のパラメータ Λ_t における条件付き確率 $p_{\Lambda_t}(s|x)$ に従ってサンプリングをする，blocked Gibbs sampling を用いて獲得したものを用いた．持橋らのアルゴリズムでは，半教師あり学習を行うために最初に教師データを用いて NPY LM の学習を行っていたが，我々の手法では教師なしで学習を行うため，NPYLM は最初に文字 n-gram のみを学習する．

5. 評価

提案手法の評価を行うため，我々は 2 つのデータセットでアルゴリズムの動作を検証した．1 つは京都大学テキス

トコーパス [*3]，もう 1 つは口語体の文章の例として，中川翔子氏の公開する「しょこたんブログ [*4]」である．

5.1 評価データ

5.1.1 京都大学テキストコーパス

京都大学テキストコーパスは，毎日新聞の 1995 年 1 月 1 日から 17 日までの全記事約 2 万文，1 月から 12 月までの社説記事約 2 万文の計約 4 万記事が含まれており，人手による正解の分かち書き及び各形態素の品詞情報が付与されている．我々はこのうちランダムで抽出した 1000 文をテストデータとし，残りを訓練データとした．

5.1.2 しょこたんブログ

しょこたんブログの解析のため，我々はブログから約 13000 記事を収集した．

実際には，しょこたんブログには顔文字等も多く含まれており，厳密に言えば口語ではない．しかし，文語体に加えて口語体も混ざって記述されているため，ここでは口語体の文章の解析のために用いた．

5.2 実験条件

京都大学テキストコーパスを用いた実験では，以下に示す 3 つの場合の分かち書きの精度を評価した．

(1) 教師なし学習
訓練データに付与されている分かち書きを削除し，文字列のみとした上で完全な教師なし学習を行う．

(2) 半教師あり学習
ランダムに抽出した 10K 文を教師データ，残りを教師なしデータとして半教師あり学習を行う．

(3) 教師あり学習
訓練データ全ての分かち書き全てを利用する教師あり学習を行う．

セグメントの最大長は 8 文字，潜在クラスのサイズ H は 20 とした．学習素性には，セグメント素性に単語 unigram 及び bigram，セグメントの長さを，Markov 素性に観測文字列の unigram，bigram，そして観測文字列の文字種 unigram，bigram，を用いた．

5.3 実験結果

5.3.1 分かち書きの精度

表 1 に，京都大学テキストコーパスでの分かち書きの評価結果を表す．

教師なしが最も低い F 値となり，半教師あり学習が最も高い F 値となった．教師あり学習は精度では最も高くなったが，再現率は半教師あり学習と比較して 5%程度低くなっている．半教師あり学習では，教師なしの事例についてはNPYLM と現在の SHDCRF のパラメータから分かち書きを獲得している．そのため，教師ありと半教師ありで学習済みのパラメータを比較すると，教師ありに対し，半教師あり学習では NPYLM に対する重みが小さく，Markov 素性の重みは大きくなっている．そのため，訓練データに現れていない未知語に対しても文字列の表層を見る Markov 素性を用いることで，適切に解析できたと考えられる．

表 1 京大コーパスでの評価
Table 1 evaluation in Kyoto Corpus

手法	精度	再現率	F 値
教師なし	0.534	0.416	0.467
半教師あり	0.862	**0.826**	**0.844**
教師あり	**0.901**	0.771	0.831

教師なし学習の結果を見ると，複合語が 1 つの形態素とされているなど，分割の基準が人手と異なる点が多く見られた．表 2 に，教師なしの分かち書き結果で得られた複合語の例を表す．これらはコーパスの中で複合語として現れることが多いため，1 形態素と見なされたと考えられる．複合語を適切に分割するには，事前知識として辞書を用いるなどが考えられるが，今回の実験ではそれは行っていない．

表 2 複合語が 1 形態素として抽出された例
Table 2 examples of noun phrases deemed as a morph

複合語の例
民主主義，連立政権，通常国会召集
細川内閣，冷戦構造，アジア・太平洋
通常国会，常任理事国，国際社会
行政改革，規制緩和，地方分権
金融機関，株式市場，信用組合
既得権益，利害関係，利益誘導
中堅議員，既成政党，国会議員
存在価値，選挙制度，解散・総選挙
村山富市首相，国際貢献，年頭所感
やさしい社会，武村正義，統一地方選
準備会参加，アジア・太平洋地域
情報通信，マルチメディア時代
非公式首脳会談，会議終了後
ボゴール宣言，韓国政府，高級事務レベル会合
人口動態統計年間推計，第二次ベビーブーム
村山富市，国会議員，毎日新聞社
海部俊樹，村山首相，羽田氏
河野洋平，クリントン政権，米露関係
北大西洋条約機構，保守派，リベラル派
軍需産業，安全保障，ポレワノフ副首相
非民営化，ポレワノフ氏，政治腐敗
ディピエトロ検事，事務局長，ヨルダン川西岸
ユダヤ人入植者，正面玄関，中央銀行

5.3.2 品詞推定

人手での品詞と学習器の与えた潜在クラスとの対応関係を，図 3 に表す．実験は，京都大学テキストコーパスの

[*3] http://nlp.ist.i.kyoto-u.ac.jp
[*4] http://ameblo.jp/nakagawa-shoko

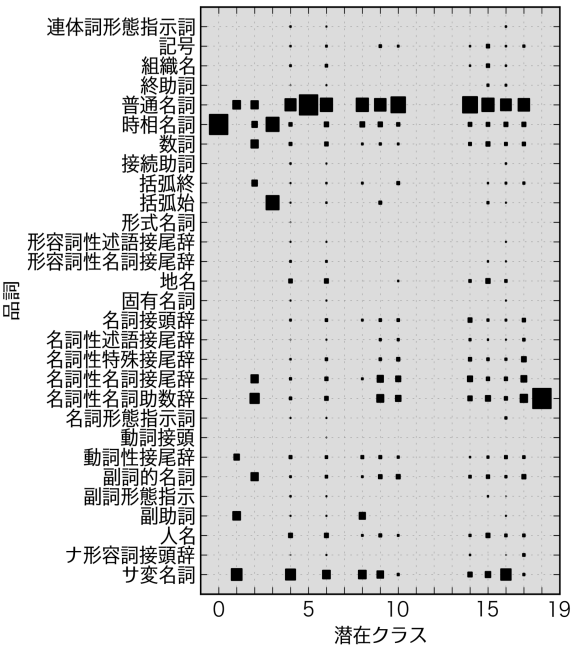

図 3　品詞と潜在クラスの対応
Fig. 3　Relations between POS and hidden variables

5000 文で学習を行った．学習データを含む全ての事例に対して形態素解析を行い，形態素の開始位置と終了位置が一致したものについて，品詞と対応する潜在クラスの分布を確認した．実験の際に用いたハイパーパラメータ α は，0.01 とした．

図 3 では，殆どの潜在クラスが普通名詞と対応付けられた．開始位置と終了位置の一致した形態素の正解の品詞の内訳を表 4 に示す．品詞の内訳を見ると，殆どが名詞であり，普通名詞とサ変名詞が多くを占める．名詞以外の品詞については分かち書きの獲得が上手くいかず，潜在クラスは名詞を細分化するように対応付けられたと考えられる．表 3 に，潜在クラスに対応する単語の例を示す．単語は，TFIDF を計算して上位 30 語を抽出した．潜在クラスの 10 を見ると，漢数字が集まっている．また，潜在クラスの 16 には，‘行’ の文字が含まれる単語が集まっている．潜在クラスの 1 から 4 については殆ど単語が割り当てられなかった．このように，一部の潜在クラスについては，表層文字列を用いてクラス分けを行う現象が見られた．

この問題を避けるためには，品詞についても半教師あり学習を行うことが挙げられる．今回の実験では，京大コーパスに付与される分かち書きのみを使用したが，利用可能な品詞も付与されているため，それを教師データとして使用することも可能であり，それによって人間の直感に沿うよう半教師あり学習で品詞推定を行うことが可能である．品詞の半教師あり学習は今後の課題としたい．

5.3.3　口語体の解析

しょこたんブログの分かち書きの例を表 5 に示す．‘#’ は文頭を，‘/’ は区切り位置をそれぞれ表す．

表 3　潜在クラスに対応する単語の例
Table 3　Examples of words corresponding hidden class

潜在クラス	単語
1	実は
2	事実, ない, な
3	、, 反対
4	都
5	こと, の, 日本, して, する, いる, で, に, など,。, これ, それ, ない, から, さ, した, は, しかし, もの, 者, 人, 問題, と, いた, 年, が, も, ため, 政府
7	の, を, が, に, は, で, と,、, ,「, 日本,。, から, 日, ・, や, も, 的な, し, でも, 年, 昨年, 経済, 国際, 一, 政治, いる, である, ため, ロシア
9	的に, 事実, に, 行わ, ない, と, 不明に, 依然と, 明確に, なかった, 出て, なければ, 行う, を, 一緒に, 強い, 見る, 持た, 入る, さ
10	「, 同, 一, 二, 三, 年, 米, 新, 四, 第, 五, 約, 国, する, 十, 大, 各, は, “, 党, と, 日, 全, 六, 今, 七, が, 核, 八
11	、,」
15	に, と, を, が, また, だが, して, で, も, さらに, ともに, 行わ, そして,、, ので, なく, のに, 確かに, した, 都, 行う, 対して, なり, まで, 滑走, か, すでに, まだ, なのに
16	、, 銀行,「, 執行,。, 現行, 旅行, 実行, 先行, ・, 犯行, 同行, 発行, 飛行機, 移行,」, 流行, 飛行, 刊行, 直行, 運行, 進行, 通行, 断行, を行っている。, 慣行, が行われた。, 施行, “
17	の,」,。, から,、, と, を, この, その, に, する, は, した, が, して, いう, いる, も, など, まで, よる, こうした, で, 中, ない, いた, ような, れた, より
18	、, を, の, に, が,「, で, 強かった,。, は, も, “, と
19	、, など, 彼ら, 大切に, 考え, 自ら, 可能に, 続く, 知ら, 違う

ブログデータは正解の分かち書きは手に入らないとの仮定のもと，教師なし学習で分かち書きの推定を行った．今回の実験では，しょこたんブログの分かち書きについて人手での正解データの作成が困難なため，定量的な評価は行えていないが，表 5 の結果から，顔文字や固有名詞，また ‘ウレシャス’ のような未知語 (造語) も教師なし学習で獲得できていることが見て取れる．

6.　まとめ

本論文では，NPYLM と SHDCRF の協調学習による教師なし・半教師あり形態素解析を提案を行った．

京都大学テキストコーパスを用いた評価により，教師なし・半教師ありで分かち書きが学習できたことを示した．また，ブログデータに対しても手法を適用し，教師なしで顔文字や造語などの未知語についても検出できることを確認した．一方で，品詞推定については潜在クラスと人間の付与した品詞は上手く対応付けられなかった．これは，実際にコーパスに含まれる品詞の粒度に差があるため，出現

表 4　品詞の内訳
Table 4　POS breakdown list

*	4746
サ変名詞	8287
ナ形容詞接頭辞	15
人名	1255
副助詞	61
副詞形態指示詞	74
副詞的名詞	345
動詞性接尾辞	611
動詞接頭辞	1
名詞形態指示詞	74
名詞性名詞助数辞	820
名詞性名詞接尾辞	672
名詞性特殊接尾辞	305
名詞性述語接尾辞	39
名詞接頭辞	283
固有名詞	47
地名	1115
形容詞性名詞接尾辞	181
形容詞性述語接尾辞	52
形式名詞	2
括弧始	59
括弧終	87
接続助詞	67
数詞	891
時相名詞	1282
普通名詞	13417
終助詞	46
組織名	276
記号	134
連体詞形態指示詞	54

頻度の大きい普通名詞に引きずられ，普通名詞を細分化するように潜在クラスが学習されたと考えられる．これに対応するため，教師データに含まれる品詞情報を用いた半教師あり学習を行うことが今後の課題である．

表 5　しょこたんブログの分かち書き
Table 5　segmentations of *syokotan-blog*

分かち書きの例
#/そして/その/あ/と/、/夏祭り/ダンスリハ/きた/お/！/初めて/の/曲/が/かなり/ある/から/がんばらないと/！ /太もも/の/内側/肉/を/ひ/き/し/め/る/方法/を/教わった/お/、/ひざ/の/間に/ＣＤ/をはさんで、/落とさないように/ずっと/テレビ/み/た/り/し/て/る/と/い/い/っ/て/！ /やるお/やるお/！ /。しようこ。/
#/マミタスさま/マミタスさま/ゴロゴロゴロゴロ/マミタスさま/(ﾟ∀ﾟ)/ただいマミタス/！！！！ /ただいマミタス/(ﾟ∀ﾟ)/マミタスさま/。しようこ。/
#/マミタスさま/と/みかん/！ /う/ま/い/かわゆす/う/ま/い/幸せ/！ /キティちゃん/みかん/！ /幸せ/(；∀；)/明日/へ/の/勇気/に/変/え/て/がんばりマミタス/！ /！ /忘れない/、/みかん色デイズ/！ /iPhoe からの投稿/
#/※/三歳/以上/は/チケット/が必要です。/
#/映画秘宝/の/お願いしたい/女優/ランキング/一位/きた/ー/恋ただ/効果/だ/と/したら/泉さん/よろこびます/よ/。しようこ/
#/さら/に/、/らいティン/の/濃厚/チーズケーキ/と/、/ルナ/の/ナイトメア/レモンタルト/！ /み/っ/つ/も/ケーキ/焼いた/！ /串田アキラさまの曲/き/いてたら/と/ま/ら/な/くなって/つい/！ /らいティン/の/チーズケーキ/は/味/見/できてない/けど/現場/もって/きた/お/！ /きょうは/夜/ま/で/ダムチャンネル収録/、/がんばる/お/！ /iPhoe からの投稿/
#/ちび太/が/かわゆす/だった/(ﾟ∀ﾟ)/階段/の/と/こ/で/かわゆす/ポーズ/で/寝た/ふりして/、/桂子に/い/じ/ってもらえるように/ずっと/待って/アピール/してるの/がたまらなく/かわゆすな/ー/しろ/い/な/ー/ちび太/か/、/レシラム/か/！ /だから/ね/、/しろ/さ/世界/の/決勝/は/！ /
#/いよいよ/マレーシア/いってき/マレーシア/(<●><●>) メポ/
#/ピンク/や/ロマンティック/や/フリル/が/だいすきな/マミタスさま/、/早速特等席に鎮座さま/、/ピンク/と/薔薇/が/お/似合い/です/さ/ま/マミタスさま/ラブマミタスさま/
#/なななななんと/モンプチ/２５周年/スペシャル/サポーター任命/式/キターーー(ﾟ∀ﾟ)/ーーーー!!!!!/キターーー(ﾟ∀ﾟ)/ーーーー!!!!!/キターーー(ﾟ∀ﾟ)/ーーーー!!!!!/キターーー(ﾟ∀ﾟ)/ーーーー!!!!!/キターーー(ﾟ∀ﾟ)/ーーーー!!!!!/２/６年/の/生涯かけて/愛/す/る/ねこ/た/ち/、/ねこ/の/ハッピー/の/ため/に/活動/できるなんて/アルティメット/ウレシャス/(ﾟ∀ﾟ)/(ﾟ∀ﾟ)(ﾟ∀ﾟ)/ーーー！！！/こんな/すごかりし/任務/、/全力でがんばります/ーーー！！！/しかも/しょこたん/デザイン/モンプチ/２５周年/セット/と/い/う/すごかりし/商品/が/できました(ﾟ∀ﾟ)/(ﾟ∀ﾟ)/(ﾟ∀ﾟ)/ーーー！！！/なんと/わたし/の/中川家寝子/８匹の/イラスト/が/エコバッグ/になりました/(ﾟ∀ﾟ)/(ﾟ∀ﾟ)(ﾟ∀ﾟ)/ーーー！！！
#/わ/！/デニーズ/が/ケルディオ/で/いっぱい/！/しらなかった/！/ケルディオ/まつり/だ/！/こ/の/ポスター/(゜д゜)ホスィ…/(ˆωˆ)/も/り/あがってる/！ /ケルディオ/！ /ポケモン/！ /iPhoe からの投稿/

参考文献

[1] 持橋大地，　鈴木潤，藤野昭典：条件付き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 (NLP2011) (2011).

[2] Shen, Y., Yan, J., Yan, S., Ji, L., Liu, N. and Chen, Z.: Sparse hidden-dynamics conditional random fields for user intent understanding, *Proceedings of the 20th international conference on World wide web*, ACM, pp. 7–16 (2011).

[3] 内元清貴，　野畑周，　山田篤，　関根聡，井佐原均：日本語話し言葉コーパスの形態素解析，言語処理学会第 9 回年次大会 (NLP2003) (2003).

[4] 松本裕治，　伝康晴：話し言葉の形態素解析，情報処理学会研究報告. 自然言語処理研究会報告，Vol. 2001, No. 54, pp. 40–54 (2001).

[5] Baroni, M., Matiasek, J. and Trost, H.: Unsupervised discovery of morphologically related words based on orthographic and semantic similarity, *Proceedings of the ACL-02 workshop on Morphological and phonological learning-Volume 6*, Association for Computational Linguistics, pp. 48–57 (2002).

[6] Argamon, S., Akiva, N., Amir, A. and Kapah, O.: Efficient unsupervised recursive word segmentation us-

ing minimum description length, *Proceedings of the 20th international conference on Computational Linguistics*, Association for Computational Linguistics, p. 1058 (2004).

[7] 松原勇介，秋葉友良，辻井潤一：最小記述長原理に基づいた日本語話し言葉の単語分割，言語処理学会第 13 回年次大会 (NLP2007) (2007).

[8] Mochihashi, D., Yamada, T. and Ueda, N.: Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, Association for Computational Linguistics, pp. 100–108 (2009).

[9] Suzuki, J. and Isozaki, H.: Semi-Supervised Sequential Labeling and Segmentation Using Giga-Word Scale Unlabeled Data., *Proceedings of ACL-08: HLT*, Association for Computational Linguistics, pp. 665–673 (2008).

[10] Sarawagi, S. and Cohen, W. W.: Semi-markov conditional random fields for information extraction, *Advances in Neural Information Processing Systems*, pp. 1185–1192 (2004).

[11] Andrew, G.: A hybrid markov/semi-markov conditional random field for sequence segmentation, *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 465–472 (2006).