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

持橋大地　鈴木潤　藤野昭典
NTT コミュニケーション科学基礎研究所
{daichi,jun,a.fujino}@cslab.kecl.ntt.co.jp

1　はじめに

日本語や中国語のように分かち書きされない言語にとって, 形態素解析は一貫して自然言語処理の主要な課題であり続けてきた. 特に最近, ブログや Twitter のようなメディアや, 今後重要性を増すと考えられる音声認識などにおいて, 従来の正書法的な新聞コーパスに基づく教師あり学習では扱えない口語的な表現や, 新語・新表現を適切に解析することがますます重要になりつつある.

これらについて, 人手で新しく多量の「正解」を準備することは難しく, そもそもこうした表現に客観的な「正解」が存在するかも疑わしい. また, 新語は有限ではなく, 口語的表現における名詞以外の未知語は連接関係が重要であるために, 辞書による対応にも限界がある. したがって, こうしたテキストの解析のためには, 適切な確率モデルの構築が不可欠といえる.

この問題に対して, 我々は先に, ベイズ学習に基づく教師なし形態素解析 (単語分割)[1]を提案した [1][2]. これは教師データを必要とせず, 観測された文字列から直接, 単語分割を学習することのできる基本的なモデルであるが[2], 言語モデルの性能を最適化しているため, 人間の分割基準と異なる場合が存在する. また, 「見 る」のように活用語尾が分離されてしまうことや, 低頻度語に弱いなどの弱点があり, 科学的意義を離れて実際的観点からは必ずしも万能ではない.

そこで本論文では, 既に存在する教師データを生かし, その基準を守りつつ, 非正書法的なテキストや口語を高精度に学習することのできる新しい半教師あり形態素解析法を提案する.

2　教師なし形態素解析: NPYLM

NPYLM (Nested Pitman-Yor Language Model) は, 我々が先に提案した教師なし形態素解析のための言語モデルおよびその学習法である. これは単語 n グラムの上に文字 n グラム (∞-グラム) モデルを持つ階層的な言語モデルであり, これ自体は言語モデルとして汎用的なものである.

NPYLM では, 概念的にはまず, 階層 Pitman-Yor 過程による文字の n グラムモデル (HPYLM) から無限個の単語とその確率が生成され, その確率分布から次に単語のユニグラム分布が生成され, そこからさらにバイグラム, トライグラム, …の分布が階層的に生成され, それを用いて実際の文字列が生成されたと考える. これは実際には, 文脈 h での単語 w の確率を計算する

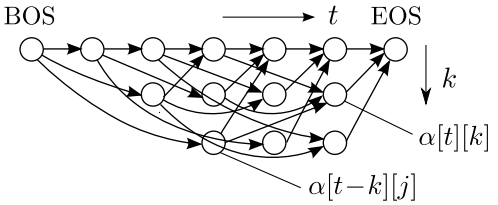


図 1: semi-Markov モデルのグラフィカルモデル (最大単語長 3 の簡単な場合). 各ノードが部分文字列に対応する.

HPYLM の再帰的な予測式 (詳細は [1] を参照)

$$p(w|h) = \frac{c(w|h) - d \cdot t_{hw}}{\theta + c(h)} + \frac{\theta + d \cdot t_{h\cdot}}{\theta + c(h)} \cdot p(w|h') \quad (1)$$

において, 1 つ短い文脈 h' が存在しない場合 (ユニグラム), バックオフ確率 $p(w|h')$ として文字 n グラム確率

$$p_0(w) = p(c_1 \cdots c_k) = \prod_{i=1}^{k} p(c_i|c_1 \cdots c_{i-1}) \quad (2)$$

を用いる階層的なスムージング法であるといってよい. これにより, 形態素解析の際に考慮するどんな部分文字列にも, 適切な非 0 の確率を与えることができる.

単語分割は非常に局所解に陥りやすいため, 学習には MCMC 法を使い, 確率的な前向き–後向きアルゴリズムを用いて各観測文字列の単語分割をサンプリングし, 言語モデルを更新していく. バイグラムの場合, この際の前向き変数 (=内側確率) は $\alpha[t][k]$ であり, これは文の時刻 t までの部分文字列 $c_1 \cdots c_t = c_1^t$ が最後の k 文字を単語として生成された周辺確率として, 次式によって計算していく.

$$\alpha[t][k] = \sum_{j=1}^{t-k} p(c_{t-k+1}^{t} | c_{t-k-j+1}^{t-k}) \cdot \alpha[t-k][j] \quad (3)$$

(3) 式は, $\alpha[t][k]$ が $\alpha[t-1][\cdot]$ から計算される通常の HMM と異なり, 図 1 のようなグラフィカルモデルを持っていることに注意されたい. これは semi-Markov HMM [5] と呼ばれるモデルであり, NPYLM は, この上できわめて精密にスムージングされた状態遷移確率 = n グラム確率を用いて, MCMC 法による前向き-後向きアルゴリズムで学習していると考えることができる.

3　識別-生成統合モデル

単語分割の生成モデルである上の NPYLM を CRF による形態素解析の識別モデルと統合し, 半教師あり学習を実現するため, 我々は半教師あり学習法として現在最高精度を持つ, 鈴木らの JESS-CM 法 [6] の枠組を採用した. JESS-CM では, 入力 $\mathbf{x}$ に対するラベル $\mathbf{y}$ の確率を次式の形で表現する.

$$p(\mathbf{y}|\mathbf{x}) \propto p_{\mathrm{DISC}}(\mathbf{y}|\mathbf{x};\Lambda)\, p_{\mathrm{GEN}}(\mathbf{y},\mathbf{x};\Theta)^{\lambda_0} \quad (4)$$

p_{DISC} は識別モデル, p_{GEN} は生成モデルであり, Θ, Λ はそれぞれのパラメータである. (4) 式は, 生成モデルに関して識別モデルが「制約」の形で重み $\lambda_0 : 1$ で働

[1]以下, "形態素解析" とは, その最も基本的な単語分割を指すものとする. 6 節最後のまとめも参照されたい.

[2]このモデルは音声認識ラティスからの単語学習 [3], 統計的機械翻訳 [4] などに最近適用されている.

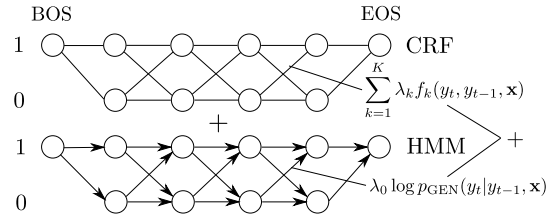


図 2: JESS-CM に基づく CRF/HMM の半教師あり学習 (本論文で扱う 2 クラスの場合).

き, その逆も成り立つことを意味している. ここで, 識別モデルを CRF のような対数線形モデル

$$p_{\mathrm{DISC}}(\mathbf{y}|\mathbf{x}) \propto \exp\left(\sum_{k=1}^{K} \lambda_k f_k(\mathbf{y}, \mathbf{x})\right) \quad (5)$$

にとれば, (4) 式は $\log p_{\mathrm{GEN}}(\mathbf{y}, \mathbf{x})$ を一つの素性関数とみて,

$$p(\mathbf{y}|\mathbf{x}) \propto \exp\left(\lambda_0 \log p_{\mathrm{GEN}}(\mathbf{y}, \mathbf{x}) + \sum_{k=1}^{K} \lambda_k f_k(\mathbf{y}, \mathbf{x})\right)$$
$$= \exp\left(\Lambda \cdot F(\mathbf{y}, \mathbf{x})\right) \quad (6)$$

と, パラメータ $\Lambda = (\lambda_0, \lambda_1, \cdots, \lambda_K)$ を持つ対数線形モデルの形でも書くことができる. ここで $F(\mathbf{y}, \mathbf{x}) = (\log p_{\mathrm{GEN}}(\mathbf{y}, \mathbf{x}), f_1(\mathbf{y}, \mathbf{x}), \cdots, f_K(\mathbf{y}, \mathbf{x}))$ とおいた. 式 (4) と (6) は等価であるから, JESS-CM ではこの二式を用いて, 教師ありデータ (X_l, Y_l) および教師なしデータ X_u からなるデータについての目的関数

$$p(X_u, Y_l|X_l; \Lambda, \Theta) = p(Y_l|X_l) \cdot p(X_u) \quad (7)$$

の値を, Λ および Θ について交互に最大化していく.

具体的には, [6] における CRF-HMM の半教師あり学習では, 図 2 のように同じ構造を持ったグラフィカルモデル上で対応するパスのコストを足し合わせ,

- Θ を固定し, $\{Y_l, X_l\}$ で CRF の重み Λ を最適化
- Λ を固定し, X_u で HMM のモデル Θ を最適化

という 2 つのステップを交互に行って (7) 式の各項を最大化してゆく. JESS-CM では, 学習の中で p_{DISC} と p_{GEN} が互いに「教え合う」ことで p_{GEN} が素性として充分賢くなり, p_{DISC} はそれをさらに教師データに合わせて補正するように学習が働く. 生成モデルの重み λ_0 は識別モデルによって自動的に決定されるが, それがまた生成モデル自身に影響を与える, という再帰的な構造になっていることに注意されたい.

4 半教師あり形態素解析: NPYCRF

上からわかるように, JESS-CM では識別モデルと生成モデルが同じ構造を持つことを前提にしている. 2 節で述べたように NPYLM は semi-Markov モデルであるから, 単語分割への自然な適用は識別モデルとして semi-Markov CRF [7] を用いることである.

しかし, これはうまく行かない. semi-Markov CRF は NE やチャンキングのように, 単語列に対して高々数語をまとめることを前提にして作られているが, 例えば日本語単語分割では文字列に対し, カタカナ語等で考慮すべき最大長が容易に 10 を超え, 空間計算量が

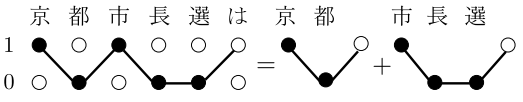


図 3: Markov→semi-Markov の重みの変換.

膨大になる.[3] さらに, semi-Markov CRF は CRF に対して精度面のアドバンテージがないことが知られており [8], 京大コーパスでの予備実験でも F 値は 95%程度であった (CRF は 99%以上).

この問題に対し, 本研究では CRF と NPYLM, すなわち 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実現する. これにより, 文字レベルの情報 (CRF) と単語レベルの情報 (言語モデル) を同時に, かつ見通しよく扱うことができる. 以下, この変換と学習アルゴリズムについて述べる.

4.1 CRF→言語モデル

Markov モデルから semi-Markov モデルへの変換は易しく, 一方向であるが [8] で試みられている. 例として図 3 を考えると, この上で「京都」→「市長選」に当たるポテンシャルは, 太線で示した経路に沿った重みを足し合わせたものになる. ここで [8] と同様に, 状態 1 は「単語の先頭」, 状態 0 は「単語の継続」を意味する. 一般に, 状態 1 で始まり 1 で終わる, 区間 $[a, b)$ $(a<b)$ の V 字型 ($b=a+1$ のときは直線) のポテンシャルを $\gamma(a, b)$ とおくと, 確率 $p(c_t^{u-1}|c_s^{t-1})$ に対応するポテンシャル $\gamma(s, t, u)$ は

$$\gamma(s, t, u) = \gamma(s, t) + \gamma(t, u) \quad (8)$$

と書くことができる. これを用いて, CRF の情報を取り入れた NPYLM の前向き確率は, (3) 式の代わりに

$$\alpha[t][k] = \sum_j \Big[\lambda_0 \log p(c_{t-k+1}^{t}|c_{t-k-j+1}^{t-k}) + \gamma(t-k-j+1, t-k, t)\Big] \cdot \alpha[t-k][j] \quad (9)$$

と計算でき, 後向きサンプリングも同様に行う.

4.2 言語モデル→CRF

一方, semi-Markov モデルから Markov モデル, すなわち言語モデルから CRF への変換は自明ではない. 例として, 図 4 において文字「京」と「都」の間にある CRF のパス (太線) を考えよう. これには 1→1, 1→0, 0→1, 0→0 の 4 通りの場合が存在する.

1→1 の場合 上の定義から, これは単語「京」の後に「都」から始まる単語が続く, すなわち

京|都, 京|都の, 京|都の法, 京|都の法案, ···

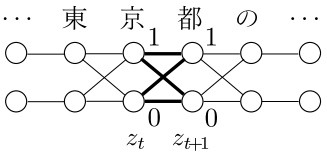
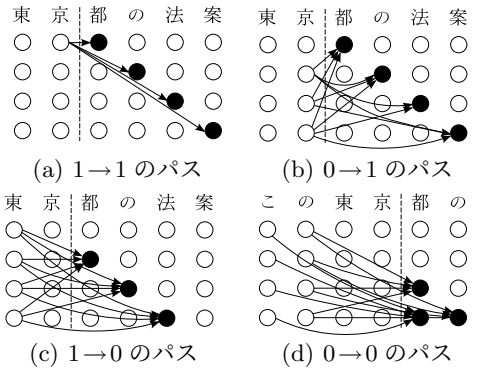


図 4: CRF の各時点での 4 通りの状態遷移.

[3] NPYLM も同じ情報を計算しているが, ラティスをメモリ上に保持する必要はない.



(a) 1→1 のパス (b) 0→1 のパス (c) 1→0 のパス (d) 0→0 のパス

図 5: semi-Markov→Markov への変換.

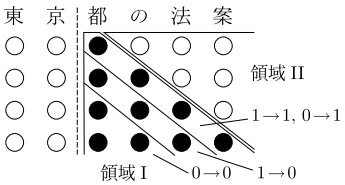


図 6: 変換に関係する終端ノード集合の構造.

を意味するから, 対応する CRF のパスの重みはこれらの言語モデル確率の和となる (図 5(a)).

0→1 の場合　これは「京」が前の単語の継続で, 次に「都」から始まる単語が続くことを意味するから, 対応する重みは

東京|都, 東京|都の, 東京|都の法, ···,
の東京|都, の東京|都の, の東京|都の法, ···,
この東京|都, この東京|都の, この東京|都の法, ···

の確率の和となる (図 5(b)).

1→0 の場合　これは「京都」から単語が始まり, 前の単語は問わないことを意味するから, 重みは

∗|京都, ∗|京都の, ∗|京都の法, ∗|京都の法案, ···

の確率の和となる. ここで, ∗ は「京都」に前接する単語すべて, すなわち「東」,「の東」,「この東」, ··· である. (図 5(c))

0→0 の場合　これは「京都」がその前からの単語の継続で, それに前接する単語は問わないことを意味するから, ∗ を同様に前接する単語群として, 重みは

∗|東京都, ∗|東京都の, ∗|東京都の法, ···,
∗|の東京都, ∗|の東京都の, ∗|の東京都の法, ···,
∗|この東京都, ∗|この東京都の, ∗|この東京都の法, ···

の確率の和となる. (図 5(d)) これは 3 重ループとなることに注意されたい.

4.2.1　図形的および数学的解釈

以上の考察は, 図 6 の形で統一的にとらえることができる. 図 5 (a)-(d) から明らかなように, 和の対象となるパスは図 6 の領域 I の黒で示したノードへ至るパスである. これらは全て, 今計算している文字「京」「都」の間を横切る. 一方で semi-Markov モデルの性質から, 領域 II へ接続するパスは決してこの間を通過せず, 和には関係しない. 図 6 に見えるように, 上記の場合分けは領域 I を左上から斜めにスライスし, その 1 行目に前接するパスの和を前列の 1 段目からと 2 段目以降からに分けてとったものが 1→1 および 0→1, 2 行目に前接するパスの和が 1→0, 3 行目以降に前接するパスの和が 0→0, という構造になっている.

数学的解釈　数学的には, これは周辺化を行っていることと等価である. まず (8) 式の定義から,

$$p(c_t^{u-1}|c_s^{t-1}) \propto \gamma(s,t,u) \tag{10}$$
$$= p(z_s{=}1,\cdots,z_t{=}1,\cdots,z_u{=}1) \tag{11}$$

であることに注意しよう. 以下, ··· の部分の変数はすべて 0 である. このとき, 求める図 4 の重みは結合確率 $p(z_t, z_{t+1})$ であり, これは確率 (11) を次のように周辺化することで求めることができる.

- $p(z_t{=}1, z_{t+1}{=}1)$
  $= \sum_k p(z_t{=}1, z_{t+1}{=}1,\cdots,z_k{=}1)$
- $p(z_t{=}0, z_{t+1}{=}1)$
  $= \sum_k \sum_l p(z_k{=}1,\cdots,z_t{=}0, z_{t+1}{=}1,\cdots,z_l{=}1)$
- $p(z_t{=}1, z_{t+1}{=}0)$
  $= \sum_k \sum_l p(z_k{=}1,\cdots,z_t{=}1, z_{t+1}{=}0,\cdots,z_l{=}1)$
- $p(z_t{=}0, z_{t+1}{=}0)$
  $= \sum_j \sum_k \sum_l p(z_j{=}1,\cdots,z_k{=}1,\cdots,$
  $z_t{=}0, z_{t+1}{=}0,\cdots,z_l{=}1)$ (12)

4.3　学習およびデコーディング

最終的な学習アルゴリズムは MCMC-EM 法に似た, 図 7 の形となる. 言語モデルの学習の際には 4.1 節で示した前向き-後向きアルゴリズムで教師なしデータ X_u の単語分割をサンプリングし, CRF の学習では言語モデルから計算した重みを 4.2 節に従って加え, 教師ありデータ Y_l, X_l について L-BFGS で最適化する.

なお, X_u でサンプルされた言語モデルから X_l での値を計算するため, このままでは計算される重みが不安定となり, 言語モデルの素性としての信頼性を示す λ_0 が不必要に小さくなるという問題が生じる. [6] ではデータ量が $|X_u| \gg |X_l|$ である上にモデルが本質的にユニグラム (HMM) のため問題は生じないが, バイグラム以上の言語モデルは非常にスパースであるために大きな問題となる. このため, 本研究では教師データの単語分割を予め言語モデルに事前知識として与えておくことにした. これは, (Y_l, X_l) と X_u を独立とする (7) 式に代えて,

$$p(X_u, Y_l|X_l;\Lambda,\Theta) = p(Y_l|X_l)\cdot p(X_u|Y_l,X_l) \tag{13}$$

と近似のない目的関数を使っていることに相当する.

デコーディング　未知データに対するデコードは, 4.2 節の重みを CRF に加えてビタビアルゴリズムを用いればよいが, 上と同様の理由で未知データに対するこの確率は値の揺れが大きく, 性能がかえって悪化する場合が多いことがわかった. 従って, 本研究では未知データに対しては, X_u の最尤分割を Y_l, X_l に加えてあらためて CRF 単体で学習したものをデコーディン

1: Add (Y_l, X_l) to NPYLM.
2: Optimize Λ on (Y_l, X_l). ⟨pure CRF⟩
3: **for** $i = 1 \cdots N$ **do**
4: **for** $j = 1 \cdots M$ **do**
5: Draw segmentations of X_u to learn Θ.
6: **end for**
7: Optimize Λ on (Y_l, X_l).
8: **end for**

図 7: NPYCRF の学習アルゴリズム.

表 1: SIGHAN Bakeoff での半教師あり学習. 辞書には libtabe[9] に含まれるものを用いた. (単位: %)

Model	CRF	NPYCRF	+辞書
Token F値	97.4	**97.5**	**97.5**
OOV 再現率	83.5	**84.1**	82.1
IV 再現率	98.5	**98.6**	**98.8**

グに用いた.[4] これは実質的に, NPYCRF を補助として教師データを「増やした」ことになる.

5 実験

日本語と中国語のブログ, CSJ 話し言葉コーパス, および中国語の単語分割テストセットを使って実験を行った. 客観評価の精密化は今後の課題である.

5.1 日本語

図 8 に, “しょこたん語”[5]として知られるスラングや特別な固有名詞の多い「しょこたんブログ」[6] の解析結果を示す (Viterbi 解). 教師データは京大コーパス 37400 文, 教師なしデータはブログから得た 40000 文である. 素性は文字および Unicode の文字種のバイグラムを用いた. 教師データに無い特有の言い回しや名詞が適切に分割されており, 多くの場合に改善されていることがわかる.

また, 図 9 は CSJ 日本語コーパスの分割を, 京大コーパスのみを教師ありデータとして行った結果である. 分割基準が CSJ の人手の基準と異なるため, 一致率の F 値は 67%程度と高くないが, 主観的にはかなり良い結果が得られている.

5.2 中国語

図 10 に, 中国語圏における Twitter である「新浪微博」[7]の学習結果を示す. 教師データは SIGHAN Bakeoff 2005 の MSR セット 87000 文, 教師なしデータは Sina API [8] を用いてランダムに取得した約 10 万文である. 新聞と異なる, 「リアル」な中国語についても, NPYCRF は適切な単語分割を与えている.

最後に, Bakeoff 2005 の MSR テストデータでの実験結果を表 1 に示す. 教師なしデータは, Chinese Gigaword の新華社通信 2004 年度分からランダムな 20 万文 (約半年分) を用いた. 精度は上昇しているが, その差は僅かである. 半教師あり学習での同じドメインでの精度向上には, テストデータをカバーするために大量の教師なしデータが必要なことが知られており [6], 計算の高速化も含めて今後の課題としたい.

なお, 素性には [9] と同じものを内製の 2 クラス CRF で用いたが, ベースライン精度の 97.4%は教師あり学習では現在世界最高値である.

やっぱり 初期 の あま あま スイート 歌い 方 の アイドル 歌謡 が すき 、 ネ申 ！最近 その ころ の 動画 発見 して ギザギザウキウキ してた お 。
ミンキー タッチ で セーラーマーズ に へーんしん ！ ！
ごろごろ ストロベリ (^ω ^) らい おっお カワユス おっお (^ω ^) (^ω ^) らす (^ω ^) ららい (^ω ^)
ぼく 小笠原 範馬 勇次郎 です

図 8: 「しょこたんブログ」の学習結果.

取り立てて 変わった 作り 方 じゃ ない んですけれども セロリ を 入れる ところが ポイント か な と 思って おり まして そして
シーツ を 被って ひゅうどろどろ と か そういった ことは でき そうに なかった ので
で この 多摩 境 な んです が 多摩 ニュータウン の 一番端

図 9: 京大コーパスのみによる CSJ の学習結果.

6 まとめ

本稿では, Markov モデルと semi-Markov モデル, すなわち CRF と言語モデルを等価変換することで可能になる, ベイズ半教師あり形態素解析を提案した. この学習法は単語分割に限らず一般的なものであり, 多クラス化により品詞推定を含めたものに拡張することは将来の課題である.

謝辞 新浪微博について正田備也氏 (長崎大) に, 中国語の素性について孫栩氏 (東大) に教えていただきました. 感謝いたします.

参考文献

[1] 持橋大地, 山田武士, 上田修功. 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.
[2] Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling. *ACL-IJCNLP 2009*, pages 100–108, 2009.
[3] Graham Neubig, Masato Mimura, Shinsuke Mori, and Tatsuya Kawahara. Learning a Language Model from Continuous Speech. *INTERSPEECH 2010*.
[4] ThuyLinh Nguyen, Stephan Vogel, and Noah A. Smith. Nonparametric Word Segmentation for Machine Translation. *COLING 2010*, pages 815–823.
[5] Kevin Murphy. Hidden semi-Markov models, 2002. http://www.cs.ubc.ca/˜murphyk/Papers/segment.pdf.
[6] Jun Suzuki and Hideki Isozaki. Semi-Supervised Sequential Labeling and Segmentation Using Giga-Word Scale Unlabeled Data. *ACL:HLT 2008*, pages 665–673, 2008.
[7] Sunita Sarawagi and William W. Cohen. Semi-Markov Conditional Random Fields for Information Extraction. *NIPS 2004*, pages 1185–1192.
[8] Galen Andrew. A Hybrid Markov/Semi-Markov Conditional Random Field for Sequence Segmentation. *EMNLP 2006*, pages 465–472.
[9] Xu Sun, Y. Zhang, T. Matsuzaki, Y. Tsuruoka, and Jun'ichi Tsujii. A Discriminative Latent Variable Chinese Segmenter with Hybrid Word/Character Information. *NAACL 2009*, pages 56–64.

站 在 我 右手 边 的 你 ， 陪伴 我 的 你 ， 同 桌 的 你 o(·"·)o
早晨 〜〜 各位 童鞋 〜〜 星期六 , 都 要 上班 〜〜
花顺 离开 了 十仔 ， 哭 死 我 了 ！ ！ ！
哥 手贱 ！ 这 下 震蛋 了 ...

図 10: 新浪微博 (中国語圏 Twitter) の学習結果.

[4] これにより X_u についても文字レベルの素性を使うことができ, 汎化性能の面でも良い影響を期待することができる.
[5] http://ja.wikipedia.org/wiki/しょこたん語
[6] http://ameblo.jp/nakagawa-shoko/
[7] http://t.sina.com.cn/
[8] http://open.t.sina.com.cn/wiki/index.php
[9] http://sourceforge.net/projects/libtabe/