テキスト分類のための単語分割

田村 晃裕　　土田 正明
日本電気株式会社
a-tamura@ah.jp.nec.com, m-tsuchida@cq.jp.nec.com

1 はじめに

テキスト分析において，テキストを所与のカテゴリに自動的に分類するテキスト分類は重要な技術であり，多くの手法が提案されている [7]．通常，テキスト分類は，まず，(i) テキストを単語分割（形態素解析）し，その後，(ii) 分割された単語に基づき素性を導出，(iii) 導出した素性を用いてモデルの学習・分類を行う[1]．単語分割は，主要言語の場合，人手による単語分割結果が付与された学習コーパスや辞書等から作成した形態素解析器により行われる．そのような資源が存在しない低資源言語においては，教師なし形態素解析 [2, 4, 8] により単語に分割できる．

これら形態素解析器や既存の教師なし形態素解析により分割された単語は，テキスト分類にとって最適な単位とは限らないという問題がある．例えば，図 1 のテキスト集合に対して,「経済」に関するテキストとそれ以外のテキストに分類することを考えると,「経済」分野を特徴付ける「国内総生産」は一単語とすべきである．しかし，形態素解析器で単語分割をする場合，学習基である学習コーパスや辞書に単語「国内総生産」が存在しないと，単語として認識してまとめることができない．特に，学習データや辞書と異なる分野のテキストを分類対象とする場合には，この問題が頻出する．一方で，教師なし形態素解析で単語分割をする場合，入力テキスト集合全体（テキスト 1〜4）で分割を最適化するため,「経済」分野に特徴的な「国内総生産」が一単語になるとは限らない．例えば，Mochihashi ら [4] の手法の場合，テキスト 1〜4 全体において単語列の尤度を最大にする言語モデルにより分割を行う．そのため,「国内」の後に「総生産」以外の文字列が出現するテキスト 3 や「総生産」の前に「国内」以外の単語が出現するテキスト 4 の影響により「国内総生産」が一単語になる確率が低くなり，一単語にまとまらない可能性が高い．

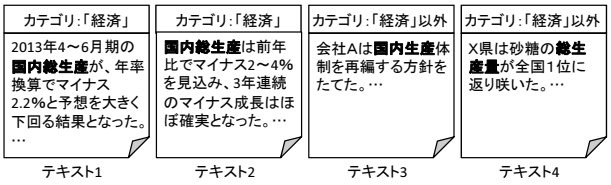

図 1: データの具体例

そこで，本稿では，分類ラベルに依存した言語モデルにより単語分割を行うことで，テキスト分類に有効な単語分割を行う教師なし手法[2] を提案する．具体的には，提案手法は，Mochihashi ら [4] の単語 n グラムモデルを分類ラベル毎に学習，適用する．ラベル未定の文の分割は，各ラベルの言語モデルを適用し，最も尤度が高い分割とする．図 1 の場合，カテゴリ「経済」と「経済以外」用の言語モデルがテキスト 1,2 及びテキスト 3,4 のそれぞれから学習されるため，カテゴリ「経済」の言語モデルにより「国内総生産」を一単語として分割することが可能になる．

以降では，2 節で提案手法のベースとなる Mochihashi ら [4] の手法を概観し，3 節で提案手法を提案する．4 節で中国語とアラビア語のテキスト分類の実験を通じて提案手法の有効性を確認し，5 節でまとめを行う．

2 従来手法：ベイズ階層言語モデル（NPYLM）による単語分割

本節では，提案手法のベースとなる Mochihashi ら [4] の手法を概観する．Mochihashi ら [4] は，教師なし単語分割の問題を，文字列 s が与えられた際に，s を分割した単語列 $\boldsymbol{w} = w_1 w_2 \dots w_N$ の確率 $p(\boldsymbol{w}|s)$ を

[1] Okanohara ら [6] や Ifrim ら [3] のように，陽に単語分割を行わないアプローチも存在するが，本研究では，様々な素性化手法との組み合わせが容易な，単語分割を一度行うアプローチに着目する．

[2] 提案手法は，単語分割において教師なし手法であり，単語分割の教師データは使用しないが，分類ラベルは利用することに注意されたい．

最大化する $\hat{\boldsymbol{w}}$ を求める問題として定式化した：

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}}\, p(\boldsymbol{w}|s). \quad (1)$$

そして，文字 n グラムと単語 n グラムをノンパラメトリックベイズ法の枠組みで統合した文字-単語の階層 n グラム言語モデル（NPYLM）を提案し，NPYLM により式 (1) 中の $p(\boldsymbol{w}|s)$ を算出した．

NPYLM では，長さ $n-1$ の文脈 $h = w_{i-n+1}\cdots w_{i-1}$ の次に出現する単語の分布（単語 n グラム分布）G_h は，$n-1$ グラム分布を基底測度とした Pitman-Yor(PY) 過程により生成される：

$$G_h \sim PY(G_{h'}, d, \theta). \quad (2)$$

式 (2) において，h' は h から一番左端を除いた文脈（$h' = w_{i-n+2}\cdots w_{i-1}$）である．また，$d$ はディスカウントパラメータ，θ は集中度パラメータである．θ は G_h が $G_{h'}$ に平均的にどのくらい似ているかを制御する．このように，単語 n グラム分布 G_h は Pitman-Yor 過程により階層的に生成される（単語 HPYLM）．

この生成プロセスは，G を積分消去した階層的な中華料理店過程で表現できる（詳細は [4] 参照）．この過程においては，単語 n グラム確率 $p(w|h)$ は次の通り階層的に計算できる：

$$p(w|h) = \frac{c(w|h) - d\cdot t_{hw}}{\theta + c(h)} + \frac{\theta + d\cdot t_{h\cdot}}{\theta + c(h)}\cdot p(w|h'). \quad (3)$$

式 (3) において，$c(w|h)$ は文脈 h で単語 w が出現した回数，t_{hw} は $c(w|h)$ のうち文脈 h' から生成されたと推定された回数，$t_{h\cdot} = \sum_w t_{hw}$，$c(h) = \sum_w c(w|h)$ である．

また，単語ユニグラム分布 G_1 の基底測度 G_0 は，単語の綴りの文字 n グラムによって与える：

$$G_0(w) = p(c_1\cdots c_k) = \prod_{i=1}^{k} p(c_i|c_1\cdots c_{i-1}). \quad (4)$$

文字 n グラム確率 $p(c_i|c_1\cdots c_{i-1})$ は，単語 HPYLM の単位を文字にした文字 HPYLM により，式 (3) と同様に算出できる．ただし，n への依存性を避けるため，文字 HPYLM では可変長の∞グラム言語モデルを用いている．このように単語 HPYLM の基底測度に文字 HPYLM が埋め込まれたモデルが，文字-単語の階層 n グラムモデル (NPYLM) である．

Algorithm 1 に NPYLM の学習アルゴリズムを示す．学習では，単語分割が付与されていない学習データ ($\boldsymbol{S}$=$\{s_1,\cdots,s_D\}$) を用いて，MCMC 法と動的計画法により，文 s_i の単語列（w(s_i)）と NPYLM（単語

Algorithm 1 NPYLM の学習アルゴリズム

1: **for** $j = 1\cdots J$ **do**
2: **for** s_i in randperm($s_1,\cdots,s_D$) **do**
3: **if** $j > 1$ **then**
4: Remove customers of w(s_i) from Θ_W and Θ_C
5: **end if**
6: Draw w(s_i) according to $p(\boldsymbol{w}|s_i, \Theta_W, \Theta_C)$
7: Add customers of w(s_i) to Θ_W and Θ_C
8: **end for**
9: Sample hyperparameters of Θ_W and Θ_C
10: **end for**

HPYLMΘ_W と文字 HPYLMΘ_C）を繰り返し改良することで，学習データに最適な単語分割と言語モデルを獲得する．具体的には，文単位のブロック化ギブスサンプリングを行う．まず，文 s_i の古い単語分割結果のデータを言語モデルから削除した後（4 行目），文 s_i の新しい単語分割 w(s_i) を $p(\boldsymbol{w}|s_i)$ に従ってサンプリングし（6 行目），新しい単語分割に基づき言語モデルを更新する（7 行目）．単語分割のサンプリングは，Forward filtering-Backwad sampling 法により効率的に行う．学習後は，未知のデータに対して，学習した NPYLM とビタビアルゴリズムにより単語分割を行うことができる．

3 提案手法 : ラベル依存 NPYLM（l-NPYLM）による単語分割

2 節の従来手法は，学習データ全体で最適化した言語モデルにより単語分割を行うため，特定ラベルに特徴的な単語が，特定ラベル以外に属するテキストの統計量の影響でまとまらない可能性がある．そこで本節では，分類ラベルに依存した言語モデル（l-NPYLM）により単語分割を行う教師無し手法を提案する．

提案手法は NPYLM を分類ラベル毎に管理し，単語列を，その文が属する文書のラベル（l）に対応する NPYLM にしたがって生成する．ただし，データスパースネスの問題を緩和するため，文字 HPYLM はラベル間で共通化する．つまり，単語 n グラム分布はラベル l 毎に独立に生成される：

$$G_{lh} \sim PY(G_{lh'}, d_l, \theta_l). \quad (5)$$

G_{lh} はラベル l に関する単語 n グラム分布である．そして，各ラベルの単語ユニグラム分布 G_{l1} の基底測度

Algorithm 2 l-NPYLM の学習アルゴリズム

1: **for** $j = 1 \cdots J$ **do**
2: **for** s_i in randperm($s_1, \cdots, s_D$) **do**
3: **if** $j > 1$ **then**
4: Remove customers of w(s_i) from $\Theta_{l_i W}$ and Θ_C
5: **end if**
6: Draw w(s_i) according to $p(\boldsymbol{w}|s_i, \Theta_{l_i W}, \Theta_C)$
7: Add customers of w(s_i) to $\Theta_{l_i W}$ and Θ_C
8: **end for**
9: Sample hyperparameters of Θ_{LW} and Θ_C
10:**end for**

G_{l0} は，ラベル共通の文字 HPYLM を用いて式 (4) により与えられる．

Algorithm 2 に l-NPYLM の学習アルゴリズムを示す．学習には，単語分割は与えられていないが文書に分類ラベルが付与された学習データ $\boldsymbol{S}$=$\{s_1, \cdots, s_D\}$ を用いる．分類ラベルの集合を L，文 s_i が属する文書のラベルを $l_i \in L$，ラベル l の単語 HPYLM を Θ_{lW}，全ての単語 HPYLM を $\Theta_{LW} = \bigcup_{l \in L} \Theta_{lW}$ で表す．Algorithm 1 と異なる部分は，NPYLM を分類ラベル毎に学習・適用して単語分割を行う点であり，Algorithm 2 中の各行の処理は Algorithm 1 の各行と同様に実現する．

l-NPYLM の学習後は，未知のデータに対して，学習した l-NPYLM とビタビアルゴリズムにより単語分割を行う．ただし，未知のデータには分類ラベルが付与されていない．そこで，各分類ラベルの言語モデルを用いて単語分割を行った後，最も尤度が高い単語分割結果 $\hat{\boldsymbol{w}}$ を採用する：

$$(\hat{\boldsymbol{w}}, \hat{l}) = \underset{\boldsymbol{w}, l}{\operatorname{argmax}}\, p(\boldsymbol{w}|s, l). \tag{6}$$

以上のように，提案手法は分類ラベル毎に単語 HPYLM を学習するため，特定ラベルに特徴的な単語を，そのラベルの単語 HPYLM により捉えることができる．

4 実験

4.1 実験設定

本節では，中国語とアラビア語のテキスト分類の実験を通じて，提案手法の有効性を検証する [3]．実験では，中国語とアラビア語のニュースサイトから収集した新聞記事に対して人手で記事の分野を付与したインハウスデータを用いた．2013 年 9 月から 2014 年 3 月までに発行された記事を使い，2013 年の記事を単語分割及びテキスト分類の学習データ，2014 年の記事をテストデータとした．評価は，記事が「経済」か否か（*ECO*）,「医療」か否か（*MED*）を分類する 2 つのタスクで行う．実験データの詳細を表 1 に示す．

言語	分野	全記事	学習	テスト
中国語	全分野	9,850	6,400	3,450
	経済	2,194	1469	725
	医療	194	134	60
アラビア語	全分野	8,374	4,472	3,902
	経済	343	227	116
	医療	135	98	37

表 1: 実験データ（記事数）

実験では，単語分割手法として，分類ラベル依存の言語モデルによる単語分割（*l-NPYLM*）に加え，ベースラインとして，分類ラベルに依存しない言語モデルによる単語分割 [4]（*NPYLM*）及び一般公開されている形態素解析器 Stanford Word Segmenter[4]（*Stanford*）を使用した際のテキスト分類性能を評価した．テキスト分類性能は，各単語分割手法で分割した結果得られた単語の Bag-of-Words を素性として，SVM[5] で学習した分類器の性能を評価した．*NPYLM* 及び *l-NPYLM* では，Gibbs iteration は 500 回行った．また，各文の単語分割における Forward filtering-Backwad sampling 法では，バイグラムにより前向き確率を計算し，単語の最大可能長 [6] は，中国語の場合は 4，アラビア語の場合は 10 とした．その他のハイパーパラメータは，Mochihashi ら [4] と同じ値にした．

4.2 実験結果

表 2 に各手法の単語分割結果を用いたテキスト分類性能を示す．参考に，陽に単語分割を行わない，Okanohara ら [6]（*all-BOW*）や Ifrim ら [3]（*SEQL*）[7] の性能も示す．*all-BOW* は，Okanohara ら [6] の手法そのものではなく，近似として，ライブラリ esaxx[8] を利用して求めた極大部分文字列を素性に使い SVM で学

[3]空白で単語が区切られていない言語では，単語分割が特に必要であるため，今回は中国語とアラビア語での評価を行った．

[4]http://nlp.stanford.edu/software/segmenter.shtml

[5]Tiny SVM （http://chasen.org/~taku/software/TinySVM/）を用いた．

[6]Mochihashi ら [4] 中の L に相当する．

[7]SEQL（http://daimi.au.dk/~ifrim/seql/seql.html）を用いた．

[8]http://code.google.com/p/esaxx/

	中国語		アラビア語	
	ECO	*MED*	*ECO*	*MED*
l-NPYLM	0.779	0.467	0.612	0.808
NPYLM	0.756*	0.453	0.595*	0.784
Stanford	0.737*	0.467	0.552*	0.784
SEQL	0.730*	0.400*	0.500*	0.622*
all-BOW	0.772	0.483	0.570*	0.757*

表 2: テキスト分類性能 (Break-even Point)

習した分類器を評価した．テキスト分類性能は，Break-even Point で評価する．Break-even Point とは，精度と再現率が等しくなる点である．また，有意差検定は，有意差水準 5%の符号検定で行う．表 2 中の「*」は提案手法 *l-NPYLM* との性能差が有意であることを示す．

表 2 より，中国語とアラビア語の両言語で，*MED* では *l-NPYLM* と *NPYLM* は同等の性能で，*ECO* では *l-NPYLM* が *NPYLM* よりも有意に分類性能が良い．この結果より，言語モデルを分類ラベルに依存させることで，テキスト分類に役立つ分割結果となり，分類性能を改善できる場合があることが実験的に確認できる．

また，中国語とアラビア語の両言語で，*ECO* の場合，*l-NPYLM* は *Stanford* よりも有意に分類性能が良い．*Stanford* の解析精度は，文献 [1, 5] によると F 値で 90%を優に超える．加えて，本実験データは新聞記事であることを考えると，本実験データに対する形態素解析精度も高いと考えられる．これより，人手で定めた単語の定義が必ずしもテキスト分類に最適とは限らず，データから最適な分割を行うことで分類性能を改善できる場合があることが実験的に分かる．

中国語の *MED* では，*l-NPYLM* は *all-BOW* より分類性能が低い．しかし，この差は有意ではないことに加えて，*all-BOW* は単語分割を行わないため，n-gram や word embedding などの BOW を超えた素性に展開しづらいことを特筆しておく．今後，BOW 以外の素性も考慮した比較を行う予定である．

5 おわりに

本稿では，分類ラベルに依存した言語モデルにより単語分割を行う手法を提案した．具体的には，Mochihashi ら [4] が提案した NPYLM の単語 HPYLM を分類ラベル毎に区別したラベル依存言語モデル *l*-NPYLM により単語分割を行う．ラベル未定の文を分割する場合，各ラベルの言語モデルによる単語分割の中で，最も尤度が高い分割とする．中国語とアラビア語のテキスト分類の実験を通じて，言語モデルを分類ラベルに依存させることにより，テキスト分類に役立つ分割となり，分類性能を改善できることを示した．

今後は，中国語とアラビア語以外の言語や，インハウスデータ以外のデータにおいても提案手法の有効性を確認する予定である．また，Mochihashi ら [4] のように，Forward filtering-Backwad sampling 法でトライグラムを使うなど，手法の改善も検討したい．

参考文献

[1] Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. Optimizing Chinese Word Segmentation for Machine Translation Performance. In *Proc. WMT 2008*, pp. 224–232, 2008.

[2] Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. Contextual Dependencies in Unsupervised Word Segmentation. In *Proc. COLING/ACL 2006*, pp. 673–680, 2006.

[3] Georgiana Ifrim, Gökhan Bakir, and Gerhard Weikum. Fast Logistic Regression for Text Categorization with Variable-length N-grams. In *Proc. SIGKDD 2008*, pp. 354–362, 2008.

[4] Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling. In *Proc. ACL-IJCNLP 2009*, pp. 100–108, 2009.

[5] Will Monroe, Spence Green, and Christopher D. Manning. Word Segmentation of Informal Arabic with Domain Adaptation. In *Proc. ACL 2014*, pp. 206–211, 2014.

[6] Daisuke Okanohara and Jun'ichi Tsujii. Text Categorization with All Substring Features. In *Proc. SDM 2009*, pp. 838–846, 2009.

[7] Fabrizio Sebastiani. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, Vol. 34, No. 1, pp. 1–47, 2002.

[8] Valentin Zhikov, Hiroya Takamura, and Manabu Okumura. An Efficient Algorithm for Unsupervised Word Segmentation with Branching Entropy and MDL. 人工知能学会論文誌, Vol. 28, No. 3, pp. 347–360, 2013.