半教師あり学習による高精度の中国語形態素解析について

王軼諞[+]、風間淳一[+]、鶴岡慶雅[*]、陳文亮[§]、張玉潔[£]、鳥澤健太郎[+]

[+]独立行政法人情報通信研究機構
[*]東京大学　[§]Institute for Infocomm Research, Singapore　[£]北京交通大学

{wangyiou, kazama,torisawa}@nict.go.jp;tsuruoka@jaist.ac.jp; wechen@i2r.a-star.edu.sg;yjzhang@bjtu.edu.cn

概要

本研究では、大規模なラベルなしデータを利用し、中国語の形態素解析精度を向上させる、いわゆる半教師あり学習に基づく手法を提案する。より具体的には、ベースラインモデルを用いて大規模ラベルなしデータを自動解析して得られる *n*-gram 情報、単語クラスタリングによって得られるクラスタ情報、交差検定法によって得られる辞書マッチング情報を追加的な素性として利用する。Penn Chinese Treebank を用いた実験では、提案手法が、半教師あり学習を用いないベースラインおよび既存手法より高い解析精度を達成することを示した。

1 はじめに

中国語には単語と単語の間に空白を入れる「分かち書き」という習慣がないため、形態素解析（単語分割と品詞タグ付け）は、中国語処理において最も基本的かつ重要な課題である。形態素解析は、構文解析や情報検索を始めとした多くのアプリケーションにおいて前処理として使用されるため、高い精度が必要である。これまで、中国語形態素解析に関して様々な研究が行われている。特に最近は、単語分割と品詞タグ付けの同時学習が多く報告されている[1,2,3,4,5]。Kruengkraiら[3]は単語-文字ハイブリッドモデルを処理方式として採用し、最高水準の解析精度を達成した。

近年、システムの性能を改善するために、正解が付与されていないデータを利用する、いわゆる半教師あり学習が盛んに使われるようになってきている。既存研究によれば、半教師あり手法を用いると、いくつかの自然言語処理タスクの性能が向上することが示されている。例えば、テキストチャンキング[6]、品詞タグ付けと固有表現抽出[7]、係り受け解析[8,9,10]などである。しかしながら、半教師あり手法を中国語形態素解析に利用した研究はあまり行われていない。持橋ら[11]は半教師あり手法で中国語の単語分割精度を向上させたが, ラベルなしデータの規模が小さいため、その差は僅かであった。

本研究では、同時学習よりも実装が容易なパイプラインシステムにおいて、大規模なラベルなしデータを利用することで、単語分割と品詞タグ付けの精度を向上させる方法を提案する。

2 形態素解析モデル

我々のシステムは、開発コストを抑えることを一つの目標とし、実装しやすい2段階のパイプラインシステムを採用している。単語分割には文字ベースのCRFを用い、品詞タグ付けには単語ベースのCRFを用いる。CRFの実装としてはCRF++（version 0.54）[1]を使用する。

2.1 ベースライン単語分割モデル

ベースラインの単語分割モデルには文字ベースの文字タグ付け法を用いる。文字タグ付け法は、単語分割の問題を、文を構成する各文字に対してその文字の単語中における位置を表すタグを付与する問題として解く。本研究では、6-タグ[12]を使用する。6-タグ(S,B,B_2,B_3,M,E)で表す単語表現を表1に、用いる素性を表2に示す。

表１　6-タグでの単語表現

単語の長さ	1	2	3	4	5	6	7
タグ	S	BE	BB_2E	BB_2B_3E	BB_2B_3ME	BB_2B_3MME	$BB_2B_3M\cdots E$

表２　単語分割の素性テンプレート

素性の説明	素性
文字 uni-gram	c_{-1}, c_0, c_1
隣接文字 bi-gram	$(c_{-1}c_0), (c_1c_0)$
ジャンプ bi-gram	(c_{-1}, c_1)
記号かどうか	$IsPu(c_0)$
文字タイプ: 時間,数字,外国語,漢字	$K(c_{-2})K(c_{-1})K(c_0)K(c_1)K(c_2)$

2.2 ベースライン品詞タグ付けモデル

我々はパイプラインシステムを採用しているため、品詞タグ付け問題は、任意の単語分割された文W＝$w_0 w_1 \cdots w_s$が与えられた時に、対応する品詞列を見つけることである。本研究では、Wuら[13]に基づいた表３に挙げる素性を用いる。

表３　品詞タグ付けの素性テンプレート

素性の説明	素性
単語 uni-gram	$w_{-2}, w_{-1}, w_0, w_1, w_2$
隣接単語 bi-gram	$(w_{-2}w_{-1}), (w_{-1}w_0), (w_1w_0), (w_1w_2)$
ジャンプ bi-gram	(w_{-1}, w_1)
単語の最初の文字	$Fc(w_0)$
単語の最後の文字	$Lc(w_0)$
単語の長さ	$Len(w_0)$

[1] http://crfpp.sourceforge.net/

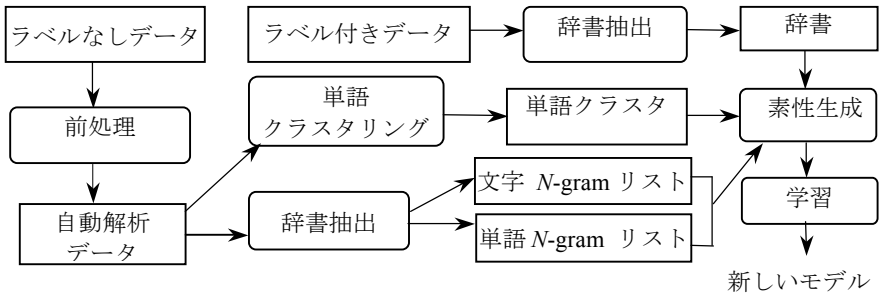

図１　提案手法の概要

3 提案手法

本節ではラベルなしデータの情報を新しい素性として前節で述べたモデルに導入するアプローチについて説明する。最初に、ベースラインモデルを用いて大規模ラベルなしデータを自動解析する。次に、自動解析データから多様な辞書情報を抽出する。そして、これらの辞書情報を単語分割と品詞タグ付けの新しい素性として利用する。さらに、分割されたデータを用い、単語クラスタリングを行い、そのクラスタ情報を品詞付けの素性として導入する。さらに、交差検定法により、ラベルありデータから抽出された辞書情報も素性に加える。本手法の概要を図１に示す。

3.1 単語分割のための新素性

(1) 半教師あり*n*-gram素性

ベースラインの単語分割モデルでラベルなしデータを単語分割し、分割された文から文字*n*-gramリストを抽出して*n*-gram素性を生成する。

ベースラインの単語分割モデルによって、ラベルなしの文の各文字c_iにタグt_iが与えられる。つまり、自動分割の結果は系列$\{(c_i,t_i)\}_{i=1}^{L}$となる。この自動分割の結果から*n*-gramリスト*{(g,seg,f(g,seg))}*が抽出される。ここで、*g*は文字*n*-gram（例えば、uni-gram c_i, bi-gram $c_i c_{i+1}$, tri-gram $c_{i-1}c_ic_{i+1}$など）を表し、*seg*は*n*-gram *g*の分割プロフィールである。分割プロフィールはタグt_iあるいはタグの組み合わせである（例えば、bi-gram $c_i c$の場合はt_iあるいは$t_i t_{i+1}$の形式で定義できる）。*f(g,seg)*は*n*-gram *g*の分割プロフィールが*seg*である時の頻度である。

そして、その頻度によって、リストを高頻度（HF：トップ5%）、中頻度（MF：5%から20%まで）と低頻度（LF：残りの80%）の３つのセットに分ける。最後に、リストL_{ng}=*{(g,seg,FL(g,seg))}*が得られる。ここで、*FL(g,seg)*は上述の方法で決めた頻度ラベルである。

n-gramリスト情報を新しい素性にエンコードするために、様々な素性表現を試したところ、*seg*=t_iのbi-gramリストから得られる素性が一番効果的であった。このリストを用い、現在の文字c_0に対して、次のように素性を生成する。L_{ng}から*g*がbi-gram c_0c_1と照合できるサブセットを獲得し、このサブセットをL_mとする。L_m中の各エントリーに対して、下記のような素性を生成する。

（a）*seg-FL(g,seg)*

そして、L_m中の各エントリーの素性を一つの*n*-gram素性として連結する。

例えば、L_mが*{(幸/福, B, HF), (幸/福,B_2, MF), (幸/福, E, LF)}*である。c_0c_1 = 「幸/福」に対して、c_0の*n*-gram素性は「*B-HF|B_2-MF|E-LF*」である。

(2) 辞書素性

文字ベースの単語分割モデルは未知語の解析精度に優れている一方、既知語の解析精度が低いことが知られている。一般的に、既知語の解析精度は辞書を用いることにより、その精度を上げることができる。既知語の辞書は、ラベルあり学習データから簡単に抽出することができる。そこで、本研究は辞書を利用することにより、素性を導入することをした。この素性を「辞書素性」という。

学習データから単語と単語に対応するすべての品詞タグを集め、辞書を作成する。例えば、「交流」に対して、エントリーの内容は(交流, *NN-VV*)である。ここで、「*NN-VV*」は学習データの中での「交流」に対応するすべての品詞タグを連結したものである。

ところが、学習データから抽出した辞書を用いて素性を生成すると、学習データへの過学習が起きる。そこで、下記の交差検定法によって、辞書を構築し、使用する。

- 学習データを10個の等しいセットに分割する。
- 各セットに対して、残りの9セットを用い、辞書を構築し、この辞書を使用し、辞書素性を生成する。

- テストセットに対しては、学習データの全体を用い、辞書を抽出し、この辞書を用いて、辞書素性を生成する。

辞書との前向き最左最長マッチを行い、単語を選ぶ。各単語wの各文字c_kに対して、下記の素性を追加する：

(b) ***P(c_k)/LEN(w)-POSs(w)***

*LEN(w)*は単語wの長さ、*P(c_k)*は文字c_kがw中の何文字目かを示す数、*POSs(w)*は単語wの辞書中の品詞タグの組み合わせを表す。例えば、文字列c_0c_1= 「幸/福」が辞書のエントリー「(幸 福, *JJ-NN-VA*)」と照合できた場合、c_0「幸」の辞書素性は「*1/2-JJ-NN-VA*」で、c_1「福」の辞書素性は「*2/2-JJ-NN-VA*」となる。

3.2 品詞タグ付けのための新素性

(1) 半教師あり*n*-gram素性

ラベルなしデータの自動分割結果は品詞タグ付けモデルで解析すると、単語レベルの*n*-gramリストL_{wg}={(w,pos,FL(w,pos))}が得られる。ここで、wは単語*n*-gramで、*pos*は単語*n*-gramの品詞プロフィールである。この*n*-gramリストを利用し品詞タグ付けの*n*-gram素性を生成する。予備実験によって、wがuni-gramで、*pos*がwの品詞である場合に、一番良い結果が得られることがわかった。L_{wg}からwが現在の単語 w_0 と照合できる照合エントリーを獲得し、このサブセットをL_sとする。例えば、w_0が「研究」である場合に、照合エントリーは(研究, *NN, HF*)、(研 究, *VV, HF*)、(研 究, *VA, LF*)と(研究, *CD, LF*)などとなる。誤り分析によって、自動タグ付けによる誤りは問題になることが多いため、サブセットL_sを獲得する際、次のように制限を設けた。*N(X)*は*FL(w,pos)=X*のエントリーの数とする。

i. *N(HF)*≧ *2*の場合は、*FL(w,pos)=HF*である照合エントリーをL_sとする。
ii. *N(HF)<2* かつ *N(HF)+N(MF)* ≧ *2* の場合は、*FL(w,pos)=HF*と*FL(w,pos)=MF*である照合エントリーをL_sとする。
iii. *N(HF)+N(MF)* <*2*の場合は、すべての照合エントリーを取る。

例えば、上記の例「研究」において、L_sは{(研究, *NN, HF*)、(研 究, *VV, HF*)}である。

単語分割と同様に、L_s中の各エントリーに対して、下記のような素性を生成する。

(c) ***pos-FL(w,pos)***

そして、L_s中の各エントリーの素性を一つの*n*-gram素性に連結する。

例えば、w_0 =「研究」に対して、w_0の*n*-gram素性は「*NN-HF|VV-HF*」である。

(2) 半教師ありクラスタ素性

自動解析のデータを用い、単語クラスタリングを行う。Kooら[10]の方法を参考にし、Brown クラスタ階層[2]のprefixを用い、様々な粒度のクラスタを作る。予備実験の結果から、下記のクラスタ素性を使用することにした。

(d) $\mathbf{w_{-1}, w_0, w_1}$の階層ビット表現の全ビット
$\mathbf{w_{-1}, w_0, w_1}$の階層ビット表現の前6ビット

予備実験では、これらのクラスタ素性をbi-gramテンプレートとして使用した場合にもっとも精度が良かった。

(3) 辞書素性

単語分割と同じ辞書を使用し、素性を追加する。現在の単語w_0に対して、下記の素性を与える。

(e) ***POSs(w_0)***

*POSs(w_0)*は辞書にある単語w_0の品詞タグを連結したものである。

4 実験

4.1 データセット

(1) ラベルありデータ

Penn Chinese Treebanksを用い、実験を行った。具体的には、CTB5(LDC2005T01)、CTB6(LDC2007T36)とCTB7(LDC2010T07)を使用した。これらのコーパスは、表4に示すように、学習セット、開発セットとテストセットに分割して用いる。既存研究ではCTB5がよく用いられるが、CTB6とCTB7ではテストセットと開発セットが増大することにより、パフォーマンスに及ぼす影響をより信頼性高く判断できる。

表4 実験用コーパス情報

	学習セットの文数	開発セットの文数	テストセットの文数
CTB5	18,089	350	348
CTB6	23,420	2,079	2,796
CTB7	31,131	10,136	10,180

(2) ラベルなしデータ

Chinese Gigaword Version 2.0(LDC2009T14) のXIN_CMN部分からCTBと重複する恐れのあるデータを取り除いて、残りの204百万語をラベルなしデータとして使用した。単語クラスタリングにはそのうち１百万語を使用した。

4.2 実験結果

提案手法の有効性を評価するために、中国語の単語分割(Seg)と品詞タグ付け(Seg&Tag)の実験を行った。精度の評価には、F値を使用した。

[2] クラスタツールはhttp://www.cs.berkeley.edu/~pliang/software/brown-cluster-1.2.zipからダウンロードできる。

表5に単語分割の実験結果を示す。*n*-gram素性と辞書素性を同時に追加する時に一番良い精度を得ている。

表5　単語分割の実験結果

単語分割方法	CTB5	CTB6	CTB7
ベースライン	0.9753	0.9513	0.9498
+(a) *n*-gram 素性	0.9798	0.9567	0.9554
+(b) 辞書素性	0.9776	0.9550	0.9542
+(a)+(b)	**0.9811**	**0.9579**	**0.9565**

表6に品詞タグ付けの結果を示す。クラスタ素性が一番効果が高く、*n*-gram素性と辞書素性と組み合わせた時に、一番良い精度が得られている。

表6　品詞タグ付けの結果

品詞タグ付け方法	CTB5	CTB6	CTB7
ベースライン	0.9362	0.9061	0.8996
+(c) *n*-gram 素性	0.9382	0.9078	0.9017
+(d) クラスタ素性	0.9403	0.9089	0.9020
+(e) 辞書素性	0.9399	0.9081	0.9019
+(c)+(d)+(e)	**0.9418**	**0.9112**	**0.9046**

4.3 実験結果の比較

表7にCTB5のデータを用いた先行研究の結果と本提案手法による結果を載せる。先行研究の結果は全て論文から引用したものである。本提案手法は単語分割も品詞タグ付けも一番良い精度を達成している。

表7　先行研究との比較（CTB5）

Method	Seg	Seg&Tag
提案手法	**0.9811**	**0.9418**
ベースライン	0.9753	0.9318
Zhangら[1]	0.9778	0.9367
Kruengkraiら[2]	0.9787	0.9367
Kruengkraiら[3]	0.9798	0.9400
Jiang ら[4]	0.9785	0.9341
Nakagawaら[5]	0.9796	0.9338

さらに、CTB6とCTB7を用い、Kruengkraiら[2]とKruengkraiら[3]に述べられている方法で、実験を行った。本提案手法による結果との比較を表8に示す。より大きいデータセットを用いて評価した場合でも本提案手法が最高精度を達成していることが分かる。

表8　先行研究との比較（CTB6とCTB7）

	CTB6		CTB7	
Methods	Seg	Seg&Tag	Seg	Seg&Tag
提案手法	**0.9579**	**0.9112**	**0.9565**	**0.9046**
ベースライン	0.9513	0.8999	0.9498	0.8937
Kruengkraiら[2]	0.9550	0.9050	0.9540	0.8986
Kruengkraiら[3]	0.9551	0.9053	0.9546	0.8990

4.4 統計学的な有意差検定

マクネマー検定を行い、解析精度の有意な改善が得られたかどうか統計学的に検証した。マクネマー検定による結果を表9に示す。CTB5におけるKruengkraiら[3]と本提案手法の間に統計学的な有意差は認められなかったが、その以外の精度差は $p < 10^{-5}$ のレベルで有意であった。

表9　マクネマー検定の結果

	p-value		
モデル	CTB5	CTB6	CTB7
提案手法 vs. [3](Seg)	0.8054	**5.0e-08**	≈ **0.0**
提案手法 vs. [3](Seg&Tag)	0.7060	**1.6e-14**	≈ **0.0**
提案手法 vs. Baseline(Seg)	**4.0e-06**	**1.8e-11**	≈ **0.0**
提案手法 vs. Baseline (Seg&Tag)	**2.1e-06**	≈ **0.0**	≈ **0.0**

5 おわりに

本稿では、パイプラインによる中国語単語分割と品詞タグ付けにおいて、簡単かつ有効な半教師あり手法を提案した。提案手法はラベルありデータを生かし、大規模なラベルなしデータから形態素情報を捉え、解析性能を向上させることができる。実験により、提案手法がベースラインおよび既存手法より高い解析精度を達成することが分かった。

参考文献

[1] Yue Zhang and Stephen Clark 2010. A Fast Decoder for Joint Word Segmentation and POS Tagging Using a Single Discriminative Model. In Proceedings of EMNLP-2010.

[2] Canasai Kruengkrai, Kiyotaka Uchimoto, Jun'ichi Kazama, Yiou Wang, Kentaro Torisawa, and Hitoshi Isahara 2009. An Error-Driven Word-Character Hybird Model for Joint Chinese Word Segmentation and POS Tagging. In Proceedings of ACL-IJCNLP-2009.

[3] Canasai Kruengkrai Kiyotaka Uchimoto, Jun'ichi Kazama, Yiou Wang, Kentaro Torisawa, and Hitoshi Isahara 2009. Joint Chinese Word Segmentation and POS Tagging Using an Error-Driven Word-Character Hybrid Model. IEICE transactions on information and systems 92(12).

[4] Wenbin Jiang, Liang Huang, Qun Liu, and Yajuan Lu. 2008. A Cascaded Linear Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging. In Proceedings of ACL-2008.

[5] Tetsuji Nakagawa and Kiyotaka Uchimoto 2007. Hybrid Approach to Word Segmentation and POS Tagging. In Proceedings of ACL Demo and Poster Sessions.

[6] Rie Kubota Ando and Tong Zhang 2005. A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. Journal of Machine Learning Research.

[7] Jun Suzuki and Hideki Isozaki 2008. Semi-Supervised Sequential Labeling and Segmentation using Gigaword Scale Unlabeled Data. In Proceedings of ACL-08: HLT.

[8] Jun Suzuki, Hideki Isozaki, Xavier Carreras, and Michael Collins 2009. An Empirical Study of Semi-supervised Structured Conditional Models for Dependency Parsing. In Proceedings of EMNLP-2009.

[9] Wenliang Chen, Jun'ichi Kazama, Kiyotaka Uchimoto, and Kentaro Torisawa 2009. Improving Dependency Parsing with Subtrees from auto-Parsed Data. In Proceedings of EMNLP-2009.

[10] Terry Koo, Xavier Carreras and Michael Collins, 2008. Simple Semi-supervised Dependency Parsing. In Proceedings of ACL-2008.

[11]持橋大地,鈴木潤,藤野昭典 2011.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. 言語処理学会第17回年次大会論文集.

[12] Hai Zhao, Chang-Ning Huang, Mu Li, and Bao-Liang Lu 2006. Effective Tag Set Selection in Chinese Word Segmentation via Conditional Random Field Modeling. In Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation.

[13]Yu-Chieh Wu Jie-Chi Yang and Yue-Shi Lee 2008. Description of the NCU Chinese Word Segmentation and Part-of-Speech Tagging for SIGHAN Bakeoff 2008. In Proceedings of the SIGHAN Workshop on Chinese Language Processing.