翻字と言語モデル投影を用いた高精度な単語分割

萩原　正人　　関根　聡
楽天株式会社楽天技術研究所
{masato.hagiwara, satoshi.b.sekine}@mail.rakuten.com

1　はじめに

日本語や中国語などの分かち書きされない言語では，単語分割および品詞付与は重要な問題の一つである．特に，カタカナ複合語は生産的であり，未知語を含むことが多いため，正確に分割することは難しい．例えば，複合語「ブラキッシュレッド」については，「シュレッド」という既知語につられ，既存の形態素解析器では「ブラキッ/シュレッド」と分割してしまう可能性が高い．しかし，逆翻字した “blacki” や “blacki shred” が英語の n グラムとして妥当ではない一方，“blacksh red” は妥当であるという知識があれば，正しい単語分割である「ブラキッシュ/レッド」を導くことができる．

中国語においては，カタカナのような翻字用の文字種が無いため問題は一層深刻である．例えば「贝拉克奥巴马」*beilake aobama* “Barack Obama” などの人名は句読点等によって分割されず，「美国总统贝拉克奥巴马在国内支持率下降」（米大統領バラク・オバマ氏の国内における支持率が低下している）という文中に直接出現する．しかし，日本語の場合と同様，「贝拉克」と「奥巴马」がそれぞれ “barack” と “obama” に逆翻字され，それが英語において妥当であるという知識を利用すれば，単語分割の際の重要な手がかりとなる．

鍛治ら [12] は，翻字と言い換えを用いて，単語境界付きのカタカナ複合語をウェブから高い精度で抽出できることを示した．しかし，この手法は単語分割用の辞書をコーパスなどからあらかじめ抽出しておく**オフライン手法**に分類され，語彙資源を更新し続けなければならないため，常に出現し続ける未知語の問題に完全に対処するのは難しい．さらに，同手法ではカタカナ語の分割のみに注目しているが，上述の通り，翻字用の文字種を持たない中国語には直接適用できない．

そこで我々は，単語分割と同時に未知語に対処する**オンライン手法** [13, 10] に注目する．具体的には，識別的構造予測モデルを拡張し，英語の逆翻字を利用し英語言語モデルの知識を日本語単語分割モデルに素性として追加（これを本稿では**言語モデル投影**と呼ぶ）し，正しい単語分割のための手がかりとして利用する．汎用の翻字モデルおよび言語モデルを用いるため，未知語に対しても頑健な解析が実現できる．著者の知る限り，翻字と英語言語モデルの知識をオンライン的手法に組み込んだのは本論文が最初である．

評価実験では，本手法を日本語書き言葉均衡コーパスおよび電子商取引 (EC) コーパス，中国語均衡コーパスにおいて評価した．その結果，両言語において単語分割精度が統計的に有意に向上した．

2　関連研究

日本語の形態素解析では，文字種等に基づく規則から未知語を生成し，品詞タグを割り当てる**未知語モデル**によってオンライン的に未知語を処理するのが一般的である．Nagata [8] は品詞，語の長さ等を考慮した未知語の生成モデルを提案しており，これら詳細な情報を考慮することにより分かち書きおよび品詞付与の精度が向上することを示した．内元ら [13] は最大エントロピー法に基づく未知語に頑健な形態素解析モデルを提案している．Peng et al. [10] は CRF の確信度の情報を使い中国語の新語を検出した．

オフライン的な未知語抽出に関しては多くの従来研究がある．Mori and Nagao [7] は未知語の周囲に出現する文字分布を既知語と比較する分布分析を用い，未知語とその品詞を獲得した．Asahara and Matsumoto [1] は，SVM を用いて文字単位のチャンキングモデルを作り，未知語を検出した．

本研究と最も近いアプローチに，鍛治ら [12] があり，ここでは，逆翻字と言い換えを用いてカタカナ複合語を分割する手法を提案している．翻字された複合語は，元の表記と併記されることが多い（例えば「ジャンクフード (junk food)」）ということや，「ジャンク・フード」などの言い換えを単語分割のための手がかりとして用いた．Nakazawa et al. [9] は，日英辞書を用いて複合語の構成要素を翻訳し，英語コーパスにおいてその出現をチェックするという類似の手法を用いている．

複合語分割の問題は，ドイツ語や韓国語など他の言語においても見られる．Koehn and Knight [4] は，単言語コーパスにおける語の統計を使い，ドイツ語複合語を分割した．また，英独対訳コーパス中に，複合語要素の翻訳が出現するかという情報も使用した．Lehal [5] は，ウルドゥー語の複合語分割のために，デーヴァナーガリー文字への翻字とヒンディー語のコーパスを用いた．

3　単語分割モデル

3.1　ベースラインモデル

本手法は識別的構造予測モデルをベースとしており，入力 x に対して，可能な解の集合 $Y(x)$ から，最適な単語分割および品詞系列 $\boldsymbol{y}^* = \arg\max_{y \in Y(x)} \boldsymbol{w} \cdot \phi(y)$ を見つける問題である．最適な解は素性 $\phi(y)$ と重みベクトル $\boldsymbol{w}$ によって決定される．

実際の単語分割は，入力の文字列を1文字ずつスキャンしながら，語の節点とそれらを結ぶ辺からなるラティス構造を構築していく．素性を $y = w_1 ... w_n$ に対して $\phi(y) = \sum_i [\phi_1(w_i) + \phi_2(w_{i-1}, w_i)]$ と分解し，最大で

ID	素性	ID	素性
1	w_i	13	$w_{i-1}^1 w_i^1$
2	t_i^1	14	$t_{i-1}^1 t_i^1$
3*	$t_i^1 t_i^2$	15*	$t_{i-1}^1 t_{i-1}^2 t_i^1 t_i^2$
4*	$t_i^1 t_i^2 t_i^3$	16*	$t_{i-1}^1 t_{i-1}^2 t_{i-1}^3 t_i^1 t_i^2 t_i^3$
5*	$t_i^1 t_i^2 t_i^5 t_i^6$	17*	$t_{i-1}^1 t_{i-1}^2 t_{i-1}^5 t_{i-1}^6 t_i^1 t_i^2 t_i^5 t_i^6$
6*	$t_i^1 t_i^2 t_i^6$	18*	$t_{i-1}^1 t_{i-1}^2 t_{i-1}^6 t_i^1 t_i^2 t_i^6$
7	$w_i t_i^1$	19	$\phi_1^{LMS}(w_i)$
8*	$w_i t_i^1 t_i^2$	20	$\phi_2^{LMS}(w_{i-1}, w_i)$
9*	$w_i t_i^1 t_i^2 t_i^3$	21	$\phi_1^{LMP}(w_i)$
10*	$w_i t_i^1 t_i^2 t_i^5 t_i^6$	22	$\phi_2^{LMP}(w_{i-1}, w_i)$
11*	$w_i t_i^1 t_i^2 t_i^6$		
12	$c(w_i) l(w_i)$		

表 1: 単語分割および品詞付与のための素性

語バイグラムの範囲に限定することにより, ビタビアルゴリズムを用いて解を効率的に求められる. 具体的には, 各 w_i に対して, 前の節点 w_{i-1} と現在の節点 w_i の間のスコアが最大化されるように辺を張る. 最後に, EOS 節点から作成した辺を逆向きに辿ることによって最適解を得る. ラティス構造の例を図 1 に, ベースラインの素性を表 1 (ID 1 ～ 18)に示す[1]. アスタリスク (*) はその素性が日本語 (JA) のみで使われ, 中国語 (ZH) には使わなかったことを示している. w_i と w_{i-1} は現在注目している語およびその前の語, t_i^j と t_{i-1}^j は対応する j 階層目の品詞である. $l(w)$ と $c(w)$ は語 w の長さと文字種の集合を表す.

3.2 未知語モデル

入力部分文字列に対して辞書見出し語が見つからない場合, 対応する節点がラティス構造に作られないため, 入力に対する解が得られない. そこで単語分割では通常, 文字種等によるヒューリスティックに基づき未知語モデルが語の候補を生成する. 日本語においては, 以下のルールに従って未知語を生成した:

- カタカナ: 連続する 1, 2, ..., 8 文字を繋げ, 全て品詞「普通名詞」を付与する.
- アルファベット: 連続する 1, 2, ..., 10 文字を繋げ, 全て品詞「普通名詞」を付与する.
- 数字: 連続する 20 文字以下の数字を繋げて単一の語とし, 品詞「数詞」を付与する.
- その他: 各文字を単一の語とし, 品詞「特殊記号」を付与する.

例えば, 図 1の入力「ブラキッシュレッド」に対して「ブ」「ブラ」「ブラキ」などが未知語節点として作成される.

中国語においては, 連続する 1, ..., 4 文字の漢字を繋げて, 「n (一般名詞)」「ns (地名)」「nr (人名)」「nz (その他固有名詞) 」の品詞を持つ 4 つの節点を作成した. また, 連続する 20 文字以下の数字を繋げて単一の語とし「m (数字)」を付与した. 「年」「月」「日」のいずれかがそれに続く場合, その文字も語の一部とし, 品詞「t (時間)」を付与した. 他の文字種に関しては, 品詞「w (その他)」の節点を作成した.

[1]本研究で用いた日本語の辞書およびコーパスの品詞体系は6階層から成るが, 中国語では1階層しかないため, 品詞素性のうちいくつかは中国語単語分割に用いなかった. 文字種としては, ひらがな (JA のみ), カタカナ (JA のみ), アルファベット, 数字, 漢字, その他, を区別した. 語長は Unicode での文字数である.

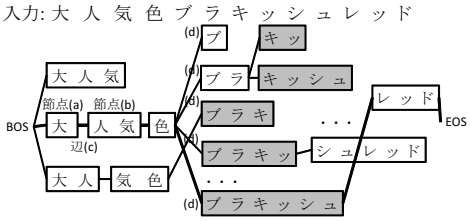

図 1: 日本語単語分割のラティス構造の例

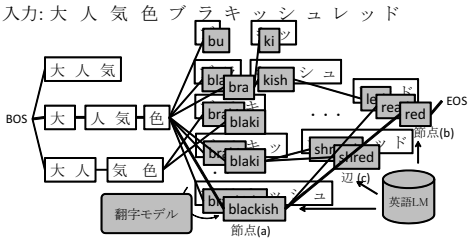

図 2: 言語モデル投影により作られるラティス構造の例

4 言語モデルの利用

4.1 言語モデル付加

言語モデル付加とは, 語 n グラムの原言語(日本語や中国語)における出現頻度を利用するものであり, n グラムの対数頻度を素性として取り入れる. これは, 表 1の素性 ID 19, 20 であり, $\phi_1^{LMS}(w_i) = \log p(w_i)$, $\phi_2^{LMS}(w_{i-1}, w_i) = \log p(w_{i-1}, w_i)$ として計算される. ここで, $p(w_i)$, $p(w_{i-1}, w_i)$ は原言語コーパスから計算される n グラム相対頻度である. もし w_i が出現しない場合, 非常に小さな確率 $p(w_i) = \varepsilon$ を仮定する. ここで, ε は最も小さい頻度の 1/10 の値である. 全ての素性値は $[0, 1]$ の範囲に正規化した. 日本語においては, 言語モデル付加はカタカナ語のみに適用した. 中国語においては, 適用範囲を制限しなかった.

4.2 言語モデル投影

言語モデル投影とは, 翻字した語 n グラムの他の言語(英語)における出現頻度を利用するものであり, 以下のようにベースラインを拡張する. まず, ラティスを構築する際に, 各節点を逆翻字し, その結果を翻字元の節点と関連付けておく. 例えば図 2 の網掛けの節点のように, 「レッド」(節点 (b))に対して英語表層形 “led”“read”“red” を持つ節点が作成される.

次に, これらの英語表層形の節点間に, 英語言語モデルの素性 (表 1 の ID 21 と 22) も考慮しながら辺を張る. 素性は $\phi_1^{LMP}(w_i) = \log p(w_i)$, $\phi_2^{LMP}(w_{i-1}, w_i) = \log p(w_{i-1}, w_i)$ として定義される. ここで, $p(w_i)$, $p(w_{i-1}, w_i)$ は英語コーパスから計算される n グラム相対頻度である. もし逆翻字に失敗したり, 英語コーパスに n グラムが出現しなかったりした場合は, 非常に小さな確率 ε を割り当てる.

最後に, 辺を EOS から辿り, 関連付けられた元の節点を単語分割の結果として使用する. 図 2 では, 太線の辺が最終的に選択された解である.

日本語においては, 翻字された語はほぼ必ずカタカナで書かれるため, カタカナ語のみを(既知・未知関わ

らず) 言語モデル投影の対象とした. 中国語においては, 表層系が 2 文字以上からなる「ns」「nr」「nz」の節点のみを対象とした. 英語言語モデルとしては, Google Web 1T 5-gram[2] を用いた.

5 翻字

逆翻字の際には任意の翻字モデルを用いることができるが, 本研究では広く用いられている Joint Source Channel (JSC) モデル [6] を用いた. JSC モデルでは, 入力語 $\boldsymbol{s}$ と出力語 $\boldsymbol{t}$ を与えたとき, それらを翻字単位 $u_i = \langle s_i, t_i \rangle$ に分割し, 翻字確率を $P_{JSC}(\langle \boldsymbol{s}, \boldsymbol{t} \rangle) = \prod_{i=1}^{f} P(u_i | u_{i-n+1}, ..., u_{i-1})$, として定義する. ここで, f は翻字単位の数である. 翻字単位とは翻字の基本となる最小単位のことであり, "la/ラ" "ish/ッシュ" のように原言語・対象言語の部分文字列のペアから構成される. 翻字単位 n グラム確率は, NEWS 2009 [3] の訓練コーパスから, EM アルゴリズムに似た逐次的更新法を用いて訓練する. 翻字候補を入力語から生成するために, 文献 [3] に示したスタックデコーダを用いた.

参考までに, 翻字そのものの性能を NEWS 2009 データを用いて評価した. 正しく翻字された語の割合は日本語で 37.9%, 中国語で 25.6% であった. この性能自体は一見低いが, 中国語から英語への逆翻字は音韻体系の違いなどから非常に難しいタスクであることに留意する必要がある.

6 評価実験

6.1 実験設定

コーパスと辞書 日本語コーパスとして (1) 汎用コーパス: BCCWJ [11] の CORE サブコーパス (60,347 文, 1,286,899 語, カタカナ語率 3.58%) および (2) ドメインコーパス: 楽天市場[4]からサンプリングした商品名・説明文 (1,230 項目, 118,355 語, カタカナ語率 11.2%) を用いた. 辞書としては, UniDic [5]を用いた.

中国語コーパスとして, Lancaster Contemporary Mandarin Corpus (LCMC) [6] (45,697 文, 1,001,549 語) を用いた. 無作為調査の結果, 翻字された語の割合は 1% 以下であった. 辞書には CC-CEDICT [7] を用いた[8].

評価指標 評価指標には, 精度 (Prec), 再現率 (Rec), および F 値を用いた. 正解コーパスに含まれる延べ語数を N_{REF}, 解析結果に含まれる延べ語数を N_{SYS}, 解析結果と正解コーパスの両者に含まれる延べ語数を N_{COR} とすると, Prec = N_{COR}/N_{SYS}, Rec = N_{COR}/N_{REF}, F = $2Prec \cdot Rec/(Prec + Rec)$ として計算される. 加えて, 複合語分割の効果を見るために, カタカナ語 (JA) もしくは固有名詞 (ZH) のみに限定した際の各指標も計算した. また, 語誤り率 (WER) も誤り率の相対変化を見るために使用した.

訓練 構造予測の重みベクトル $\boldsymbol{w}$ の学習には平均化パーセプトロンを用い, 性能は 5 分割交差検定により評価した. なお, 重みを更新するためには, 訓練文の素性値が必要となる. 訓練コーパス中の各語に対する正しい逆翻字結果は分からないため, 英語言語モデルにおいて最も高い確率を持つ語を翻字結果とし, 素性値 ϕ_1^{LMP} および ϕ_2^{LMP} を計算した.

6.2 結果

表 2上段に, BCCWJ における単語分割結果を示した. ここでは, ベースライン, 言語モデル付加 (+LM-S), 言語モデル投影 (+LM-P), ならびに各種オープンソースの日本語形態素解析システムを比較している[9]. +LM-S では性能はほとんど変化していないが, +LM-P では, F 値の値が 0.03 ポイント向上した. この差異は非常に小さいように思えるが, コーパスが大きいこと, カタカナ語の割合が高くない (約 3.58%) こと, ベースライン性能が既に高いことなどを考えると, 影響は大きく, 実際に統計的に有意であった. 実際, カタカナ語のみの F 値は約 1 ポイント上昇し, WER も 1 パーセント減少している. JUMAN のカタカナ語分割性能が高いが, これは JUMAN が Wikipedia やウェブなどから獲得した大規模語彙集合を備えているためである.

提案手法の優位性は, カタカナ語の割合が比較的高い EC ドメインコーパスにおいて評価した場合さらに顕著であった (表 2中段). ここでは, F 値が 0.48 ポイント上昇し, WER が 16.0% 減少した. MeCab+UniDic がわずかながら高いカタカナ語分割性能を示したが, これはより大きい訓練コーパス (BCCWJ) において訓練されているためであると考えられる.

+LM-S による改善の多くが, 「レイン/スーツ」「ルーカス/フィルム」など, より粒度の細かい分割による. 実際, +LM-S と +LM-P との差異を 30 個調査したところ, そのうち 25 個 (83%) において真の改善が見られた. 一方で, 「キーワード」を「キー/ワード」として過剰に分割してしまうという現象も見られた. これは正解とは異なるが, 情報検索や機械翻訳では, より粒度の細かい分かち書きが有効であるという知見 [2] もあるため, 悪影響があるとは言えない. また, 「スノコ/ベッド」などの和語由来の複合語や, 「ガソリン/スタンド」などの和製英語に弱いという特徴も見られた.

表 2下段に, 中国語における単語分割結果を示した. CRF に基づく中国語単語分割器 Stanford Chinese word segmenter [10] の結果も比較のため示した. ここでは, +LM-S によって性能が低下していることが分かるが, これは日本語のように適用範囲をカタカナ語に制限できないことによる副作用だと考えられる. 一方, +LM-S を +LM-P によって置き換えることにより全体の性能は向上しており, 「欧麦尔/萨利赫」*oumaier/salihe* "Umar Saleh" や「领导人/曼德拉」*lingdaoren/mandela* "Leader Mandela" などを正しく分割できた. ただ, このコーパス

[2] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2006T13
[3] http://www.acl-ijcnlp-2009.org/workshops/NEWS2009/index.html
[4] http://www.rakuten.co.jp/
[5] http://www.tokuteicorpus.jp/dist/
[6] http://www.lancs.ac.uk/fass/projects/corpus/LCMC/
[7] http://www.mdbg.net/chindict/chindict.php?page=cedict
[8] CC-CEDICT には, 品詞が明示的に記載されていないため, 訓練コーパスに現れる見出し語のみに限定し, 現れた品詞タグを可能な品詞として付与した.

[9] MeCab+UniDic および KyTea は BCCWJ で訓練されているため, この比較は単に参考用である.
[10] http://nlp.stanford.edu/software/segmenter.shtml なお, Stanford segmenter の単語分割基準は我々のものとは異なり, また, 本実験では訓練セットと CC-CEDICT との共通集合しか用いていないため, これは公平な比較とは言えない. Stanford segmenter は品詞を付与しないため, 固有名詞のみの評価は行なっていない.

コーパス	モデル	Prec. (O)	Rec. (O)	F (O)	Prec. (P)	Rec. (P)	F (P)	WER
日本語 BCCWJ	MeCab+IPADic	91.28	89.87	90.57	88.74	82.32	85.41	12.87
	MeCab+UniDic*	(98.84)	(99.33)	(99.08)	(96.51)	(97.34)	(96.92)	(1.31)
	JUMAN	85.66	78.15	81.73	**91.68**	**88.41**	**90.01**	23.49
	KyTea*	(81.84)	(90.12)	(85.78)	(99.57)	(99.73)	(99.65)	(20.02)
	ベースライン	96.36	96.57	96.47	84.83	84.36	84.59	4.54
	+LM-S	96.36	96.57	96.47	84.81	84.36	84.59	4.54
	+LM-S+LM-P	**96.39**	**96.61**	**96.50**	85.59	85.40	85.50	**4.50**
日本語 EC	MeCab+IPADic	84.36	87.31	85.81	86.65	73.47	79.52	20.34
	MeCab+UniDic	95.14	**97.55**	96.33	**93.88**	**93.22**	**93.55**	5.46
	JUMAN	90.99	87.13	89.2	92.37	88.02	90.14	14.56
	KyTea	82.00	86.53	84.21	93.47	90.32	91.87	21.90
	ベースライン	97.50	97.00	97.25	89.61	85.40	87.45	3.56
	+LM-S	97.79	97.37	97.58	92.58	88.99	90.75	3.17
	+LM-S+LM-P	**97.90**	**97.55**	**97.73**	93.62	90.64	92.10	**2.99**
中国語 LCMC	Stanford Segmenter	87.06	86.38	86.72	—	—	—	17.45
	ベースライン	90.65	90.87	90.76	**83.29**	51.45	**63.61**	12.21
	+LM-S	90.54	90.78	90.66	72.69	43.28	54.25	12.32
	+LM-P	**90.90**	**91.48**	**91.19**	75.04	**52.11**	61.51	**11.90**

表 2: 単語分割性能 (%) — 日本語: 全体 (O) およびカタカナ (P), 中国語: 全体 (O) および固有名詞 (P)

における翻字語率の低さ(1% 以下)を考えると, 言語モデル投影による F 値の向上は単に翻字語のみによるものではなく, 全体の分割粒度をより小さくする方向へと寄与していることが示唆される.

中国語の言語モデル投影が特に難しい理由の一つに, 今回用いたコーパスでは,「马克思主义者」*makesizhuyizhe*(マルクス主義者)(马克思 *makesi*「マルクス」+ 主义者 *zhuyizhe*「主義者」) のように翻字された語に接辞が付与されている形態素がある. 他にも,「尼罗河」*niluohe*「ナイル川」(尼罗 *niluo*「ナイル」+ 河 *he*「川」) のような例がある. また, 翻字が誤りの原因となった場合もある. 例えば,「维娜斯」*weinasi*「ヴィーナス」に対して適切な翻字が生成されなかったが, これは「维纳斯」*weinasi* と翻字されるのが一般的である. 翻字モデルを改善することにより, 言語モデル投影の性能を改善することができる可能性がある.

7 おわりに

本稿では, 日本語および中国語における複合語分割の問題に対処するために, 逆翻字と英語の言語モデルを統合したオンライン単語分割手法を提案した. 実験により, 提案手法は統計的に有意な向上をもたらし, 特に EC サイト分野のコーパスにおいて語誤り率を 16% 低下させることができることが分かった.

提案手法の問題点の一つは速度であるが, 逆翻字モデルの手法にスタックデコーダを用いていることから, 一つ前の結果を再利用することにより高速化が可能である. また, 言語モデル投影という概念の適用範囲はカタカナ語や中国語の翻字語に留まらない. 例えば,「仲里依紗」という人名を姓と名に分割する際, *Nakariisa* という読みを推定し, 英語の言語モデルにおける *Naka Riisa* というユニグラムおよびバイグラムの頻度を手がかりにすることにより,「仲/里依紗」と正しく分割できると考えられ, 今後さらに適用範囲を検討する予定である.

参考文献

[1] Masayuki Asahara and Yuji Matsumoto. Japanese unknown word identification by character-based chunking. In *Proc. of COLING*, pp. 459–465, 2004.

[2] Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. Optimizing chinese word segmentation for machine translation performance. In *Proc. of WMT*, pp. 224–232, 2008.

[3] Masato Hagiwara and Satoshi Sekine. Latent class transliteration based on source language origin. In *Proc. of NEWS*, pp. 30–37, 2012.

[4] Philipp Koehn and Kevin Knight. Empirical methods for compound splitting. In *Proc. of EACL*, pp. 187–193, 2003.

[5] Gurpreet Singh Lehal. A word segmentation system for handling space omission problem in urdu script. In *Proc. of WSSANLP*, pp. 43–50, 2010.

[6] Haizhou Li, Zhang Min, and Su Jian. A joint source-channel model for machine transliteration. In *Proc. of ACL*, pp. 159–166, 2004.

[7] Shinsuke Mori and Makoto Nagao. Word extraction from corpora and its part-of-speech estimation using distributional analysis. In *Proc. of COLING*, pp. 1119–1122, 1996.

[8] Masaaki Nagata. A part of speech estimation method for japanese unknown words using a statistical model of morphology and context. In *Proc. of ACL*, pp. 277–284, 1999.

[9] Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi. Automatic acquisition of basic katakana lexicon from a given corpus. In *Proc. of IJCNLP*, pp. 682–693, 2005.

[10] Fuchun Peng, Fangfang Feng, and Andrew McCallum. Chinese segmentation and new word detection using conditional random fields. In *Proc. of COLING*, 2004.

[11] 前川喜久雄. Kotonoha『現代日本語書き言葉均衡コーパス』の開発. 日本語の研究, Vol. 4, No. 1, pp. 82–95, 2008.

[12] 鍛治伸裕, 喜連川優. 言い換えと逆翻字を用いた片仮名複合名詞の分割. 自然言語処理, pp. 65–88, 2012.

[13] 内元清貴, 関根聡, 井佐原均. 最大エントロピーモデルに基づく形態素解析 — 未知語の問題の解決策 —. 自然言語処理, Vol. 8, pp. 127–141, 2001.