日本語形態素解析辞書の英訳候補付与とその応用

山本 薫 †　　奥村 学 ‡
†CREST JST　　‡ 東京工業大学 精密工学研究所
kaoru@lr.pi.titech.ac.jp　　oku@pi.titech.ac.jp

1 借用語と日本語形態素解析

「インパクト (impact)」のような外来語から「衝撃」のような日本語へ言い換える提案が、国立国語研究所によってなされている[1]。的確な言い換え表現が定着するまでには時間がかかる。それまで、我々は、借用語（翻訳借入語、calque）をカタカナで表記し、利用する。マニュアル、学術論文、特許など特殊文書では、カタカナ表記した借用語が目立つ。

JUMAN や ChaSen などの日本語形態素解析システムでは、多くの借用語は未知語である。未知語処理は、オンラインで、わかち書きと品詞付与を行なう。わかち書きされていない分、未知語の範囲を適切に分割することが困難である。通常は、オフラインで、新語を獲得し、辞書に登録して、未知語処理の負担を減らす戦略がとられる。どのように、新語を形態素解析辞書に追加し、頑健な解析を実現するかは、重要な課題である。

1990 年代より、対訳コーパスから対訳を抽出する方法が提案され、最近では、Web から新しい対訳文や対訳表現を収集する手法が報告されている。同時に、電子辞書も普及し、専門分野に特化した対訳辞書が、入手できるようになった。

対訳表現の多くは、連続文字列から成り立つ。新語として、形態素解析辞書に追加できるものが多い。見出し語として登録すれば、訳出単位を尊重したわかち書きが実現でき、対訳単語アライメントには好都合である。さらに、対応する英訳候補も登録しておけば、形態素レベルでの翻訳が可能になる。内山は、この仕組みを利用して、ChaSen の内部辞書 IPADIC に EDR など日本語側の見出し語を追加し、日本語記事の英単語集合への変換を行なっている [1]。新語のわかち書きを補正するために、対訳辞書を活用した。

新語の品詞付与にも、対訳辞書が使えないだろうか。直観的には、借用語の品詞は、由来元の接頭や接尾、品詞と関連がありそうである。例えば、「ロシア」や「サッチャー」など固有名詞は、大文字始まりで表記される。「コーパス」は、英語では名詞のみなので、用法は普通名詞に近いが、「アドバイス」は、英語では名詞にも動詞にもなれるので、サ変名詞になり、「ユニーク」は、英語の品詞がそのままで使われ、ナ形容詞（形容動詞）となる。ここまで考えると、借用語の品詞は、規則で付与できそうである。だが、「コミュニケーション」のように、英語では名詞用法のみだが、日本語ではサ変名詞としてふるまう[2]といった例外もあり、一筋縄では行かない。

本稿では、借用語の品詞を適切に付与する問題を扱う。新たに対訳辞書から得られる属性をも取り込める形態素解析モデルを試作した。2 節で、手順を述べ、3 節で、実験結果を報告する。見出し語の文字列を完全に一致させ、英訳と品詞の候補を探し、対訳素性を展開して、形態素解析モデルに取り入れても、あまり効果がないことがわかった。

2 借用語の品詞付与

2.1 Conditional Random Fields

形態素解析モデルは、工藤の手法 [2] を踏襲して、Conditional Random Fields (CRFs) で学習する。CRFs は、重なりあう素性を許すため、由来元の接頭や接尾、品詞などが考慮できる枠組である。一般的に、対訳辞書のわかち書き単位は形態素辞書の単位より長い。そのため、見出し語を追加しただけでは、分割数が小さい（極端に曖昧性が小さい）系列が誤って選ばれやすい。CRFs は、このバイアスを抑制し、正しい品詞列を推定するのに優れている。

CRFs を用いた日本語形態素解析では、ある出力系列 $\mathbf{y} = (\langle w_1, t_1\rangle \ldots \langle w_{\#\mathbf{y}}, t_{\#\mathbf{y}}\rangle)$ の入力文 $\mathbf{x}$ に対する条件付き確率 $P(\mathbf{y}|\mathbf{x})$ をモデルとする。

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_{\mathbf{x}}} \exp \sum_{i=1}^{\#\mathbf{y}} \sum_{k} \lambda_k f_k(\langle w_{i-1}, t_{i-1}\rangle, \langle w_i, t_i\rangle) \ ,$$

$$Z_{\mathbf{x}} = \sum_{\mathbf{y}' \in \mathcal{Y}(\mathbf{x})} \exp \sum_{i=1}^{\#\mathbf{y}'} \sum_{k} \lambda_k f_k(\langle w'_{i-1}, t'_{i-1}\rangle, \langle w'_i, t'_i\rangle) \ ,$$

$f_k(\langle w_{i-1}, t_{i-1}\rangle, \langle w_i, t_i\rangle)$ は i 番目と $i-1$ 番目の出力ラベルに依存する素性関数である。ここでは、HMM での出力確率に相当する singleton 素性関数と遷移

[1] http://www.kokken.go.jp/public/gairaigo/Word_List/iikaego.html

[2] Google で、”コミュニケートする ”の検索結果は約 14,400 件だが、”コミュニケーションする ”は約 44,700 もあり、後者も一般的に用いられていると推測した。

表 1: 対訳属性

属性	説明	
jpos	（日）品詞	
jsub	（日）細分類	
jcty	（日）活用型	
jcfm	（日）活用形	
jbse	（日）基本形	
jchr	（日）文字種	
epre	（英）接頭辞	大文字か小文字
esuf	（英）接尾辞	ion, ing, ent, nce, ed, er, al
epos	（英）品詞集合	名詞、動詞、形容詞、副詞

確率に相当する doubleton 素性関数だけを考慮する。$\lambda_k (\in \mathbf{\Lambda} = \{\lambda_1, \ldots, \lambda_K\} \in \Re^K)$ は、素性関数 f_k の重みである。$Z_{\mathbf{x}}$ は、正規化項である。

入力 $\mathbf{x}$ に対する最適な出力 $\hat{\mathbf{y}}$ は、次の式で求める。

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} P(\mathbf{y}|\mathbf{x}) = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} \frac{1}{Z_{\mathbf{x}}} \exp(\mathbf{\Lambda} \cdot \mathbf{F}(\mathbf{y}, \mathbf{x})) = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} \mathbf{\Lambda} \cdot \mathbf{F}(\mathbf{y}, \mathbf{x}),$$

$\mathbf{F}(\mathbf{y}, \mathbf{x}) = \{F_1(\mathbf{y}, \mathbf{x}), \ldots, F_K(\mathbf{y}, \mathbf{x})\}$ は大域ベクトルで、$F_k(\mathbf{y}, \mathbf{x}) = \sum_{i=1}^{\#\mathbf{y}} f_k(\langle w_{i-1}, t_{i-1} \rangle, \langle w_i, t_i \rangle)$ とする。構築されたラティス内の最適解は、素性関数 f_k とその重みである λ_k から得た系列コストを使って、Viterbi アルゴリズムで探索する。

2.2 対訳関係を考慮した素性関数

借用語は、原則として、名詞になる [3]。JUMAN や ChaSen は、未登録の借用語（カタカナ文字列）をサ変名詞（もしくは普通名詞）とみなし、未知語と出力する。本稿のねらいは、借用語の品詞が、未知語ではなく、適切に付与されることである。

CRFs の登場により、対訳関係を素性関数（対訳素性と呼ぶ）として盛り込めるようになった。以下では、新規に投入した対訳素性について述べる。日本語のみから構成される素性テンプレートは、従来通り、工藤 [2] と同じものを利用した。

まず、属性の表記について説明する。素性テンプレートを $f_k(\langle w', t' \rangle, \langle w, t \rangle)$ とする。w と t の属性を表 1 にまとめる。英語の接頭辞 (epre) は、英訳候補の中に、最初の一文字が大文字で始まる訳語があるかを表現した。固有名詞の訳語は大文字始まりが多いという特徴をつかむためである。英語の接尾辞 (esuf) は、英訳候補の中に、頻出語尾で終了する訳語があるかを表現した。英語の未知語処理で、接尾辞を品詞推定に利用する報告がきっかけになった。英語の品詞候補集合 (epos) は、英訳候補が、自立語の品詞になりうるかどうかを表現した。英訳が名詞用法だけなら普通名詞になりそうだが、動詞用法もあればサ変名詞になりそう、といった直観を素性関数として表現するためである。

表 2: 対訳素性: singleton(上) doubleton (下)

	説明	対応する属性
接頭 s_11	大文字始まり	jpos, jsub, jfrm, epre
接頭 s_12		jpos, jsub, jfrm, jchr, epre
接尾 s_21	頻出語尾	jpos, jsub, jfrm, esuf
接尾 s_22		jpos, jsub, jfrm, esuf, epos
品詞 s_31	品詞の対応	jpos, epos
品詞 s_32		jpos, jsub, jfrm, epos
	前件	後件
接頭 dl_11	接頭 s_11’	jpos, jsub, jcfm
接頭 dl_21	接尾 s_21’	jpos, jsub, jcfm
接尾 dl_22	接尾 s_22’	jpos, jsub, jcfm
品詞 dl_31	品詞 s_31’	jpos, jsub, jcfm
品詞 dl_32	品詞 s_32’	jpos, jsub, jcfm
接頭 dr_11	jpos’, jsub’, jcfm’	接頭 s_11
接頭 dr_21	jpos’, jsub’, jcfm’	接尾 s_21
接尾 dr_22	jpos’, jsub’, jcfm’	接尾 s_22
品詞 dr_31	jpos’, jsub’, jcfm’	品詞 s_31
品詞 dr_32	jpos’, jsub’, jcfm’	品詞 s_32

これらの属性から作成した singleton 素性テンプレートを表 2 に示す。それぞれ、日本語と英語の属性を組み合わせて、実現した。

借用語の周辺にも特徴的な連接が観察される。例えば、後件の基本形が「〜する」のようなら前件の借用語はサ変名詞、同様に、「〜だ」ならナ形容詞と推測できる。「両サイド」のように、前件が名詞接頭辞があれば、後件の借用語は普通名詞になりやすそうである。

品詞の連接を明示的にとらえるために、doubleton 素性テンプレートも設計した。前件か後件のいずれかが、普通名詞、サ変名詞、固有名詞（人名、地名、組織名を含む）、ナ形容詞のどれかであれば、doubleton 素性関数を適用する。

2.3 英訳候補の付与

ここまでは、学習に使う辞書やコーパスから、対訳素性が抽出できることを仮定して、議論を進めてきた。しかし、日本語の形態素辞書や品詞タグ付きコーパスには、英訳語及び英品詞は付与されていない。この節では、英訳候補の自動付与手順について述べる。

本稿では、JUMAN 辞書、京大コーパス、EDICT、Penn Treebank、Suzanne Corpus を用いた。以下では、これらの言語資源について述べるが、類似資源の組み合わせでも代用できる。

はじめに、JUMAN 品詞全体から、訳出される品詞と訳出されない品詞へ分類した。訳出される品詞は、名詞（普通名詞、サ変名詞、固有名詞、地名、人名、組織名、時相名詞）、動詞、形容詞、副詞とした。

次に、訳出される品詞を持つ見出し語に訳語を付

表 3: 英訳候補の自動付与

		辞書			コーパス		
品詞	細分類	日本語	英訳語	英品詞	日本語	英訳語	英品詞
名詞	普通名詞	21583	11504	6200	8580	5243	3453
	サ変名詞	6559	4220	2973	2863	2186	1703
	固有名詞	11	7	7	10	4	4
	地名	9241	4435	4435	984	617	617
	人名	20126	12159	12159	1950	1099	1099
	組織名	433	104	104	616	113	113
	時相名詞	412	304	298	241	189	184
動詞		14066	4346	4235	4213	3203	3175
形容詞		8401	3009	2571	1511	1244	1139
副詞		2015	1017	970	560	361	352

与した。訳語は、JUMAN の見出し語と EDICT の日本語の見出し語の文字列一致により検索した。ただし、活用する品詞（動詞、形容詞）は、見出し語のみならず、基本形や語幹で照合させた。

最後に、訳語に英語品詞を付与した。あらかじめ、英語の品詞タグ付きコーパス Penn Treebank と Suzanne Corpus から、英語の見出し語がとりうる品詞集合を列挙した。品詞は、訳語と英語コーパスの見出し語の文字列一致により、とりうる品詞集合を検索した。日本語の品詞が、固有名詞、地名、人名、組織名の場合は、英語の品詞検索を行なわず、名詞とした。訳語が複数語の場合は、英語処理で使われるストップワードを削除して残った末尾語の品詞集合を検索した。対訳属性 epos は、訳語が、名詞、動詞、形容詞、副詞になりうるかを抽出するので、訳語の品詞は、可能性に基づく品詞集合として表現した。

同様の手順で、京大コーパスにも訳語と品詞を付与した。自動付与の結果を表 3 に示す。JUMAN 辞書は、バージョン 4.0 を用い、異表記同語のうち、重みが 1 の見出し語を訳語付与の対象にした。京大コーパスは、バージョン 3.0 を用い、１月１日から１月９日までの記事を対象にした。表 3 の統計は、辞書に関しては、異表記同語を別々に展開して数え、コーパスに関しては、形態素を見出し語の異なりで数えた。自動付与なので、誤りも含まれるが、約半分の英訳候補が付与されている。

3 実験と結果

3.1 実験

借用語は、京大コーパスの 1 月 9 日付の記事のみに出現したカタカナ文字列とした。計 328 形態素になる。内訳を表 4 に示す。日本語の見出し語に、英語の訳語と品詞を付与した。明らかな誤りは、人手で修正した。これを追加する新語辞書とする。

日本語形態素辞書辞書に新語を登録するとき、新語の品詞が明白な場合と曖昧な場合がある。そこで、次の 3 つのシナリオを想定した。理想的なシナリオでは、新語の品詞は明白で、正解がわかると仮定した。京大コーパスで付与されている品詞を使った。現実的なシナリオでは、新語の品詞は曖昧で、名詞か形容詞になりうると仮定した。2 つの品詞に付随するすべての細分類、活用型、活用形を列挙した。基本的なシナリオでは、新語の品詞は支配的な品詞と仮定した。表 4 を参考にして、すべて、普通名詞とした。

表 4: 借用語: 328 形態素の内訳

品詞	細分類	例	数
名詞	普通名詞	アーケード	148
名詞	サ変名詞	コピー	32
名詞	地名	フロリダ	28
名詞	人名	サダエフ	76
名詞	組織名	グリーンピース	28
形容詞		アヴァンギャルド	6
副詞		グッ	8
感動詞		グーテンターク	2

実験では、基本辞書にそれぞれのシナリオで新規辞書を追加し、解析モデルを学習し、全形態素及び借用語の品詞付与精度を比較する。基本辞書とは、JUMAN 辞書から新語辞書と重複した 119 エントリを削除した辞書を指す。学習データは、1 月 1 日から 8 日付の記事を利用した。延べ 189804 形態素ある。評価データは、1 月 9 日付の記事を用いた。延べ 30607 形態素ある。

CRFs の実装及びパラメータ学習は、工藤の手法 [2] に従った。詳細は、文献に譲る。過学習を防ぐために、パラメータを Gaussian Prior で正規化し、そのハイパーパラメータ C は、交差検定より 1.1 に設定した。評価では、わかち書き及び品詞情報（品詞、細分類、活用型、活用形）がすべて一致していれば、正解とした。F 値（$F_{\beta=1}$）で、それぞれの結果を比較する。

表 5: 実験結果

		理想		現実		基本		mzc–	jmn–	jmn+
対訳素性	形態素数	×	○	×	○	×	○	n/a	n/a	n/a
全体	(30607)	95.964	95.849	95.069	94.987	95.384	95.229	94.537	92.508	93.896
カタカナ	(1246)	88.237	87.836	66.104	66.746	73.253	72.610	58.893	50.791	84.273
新語	(396)	96.534	96.287	42.661	48.872	65.442	65.108	31.063	0.503	96.388

表 6: 対訳素性のサンプル

重み	テンプレート	素性
1.89869091906657061	接頭 s_21	大文字_カタカナ_名詞_人名
0.97944732741720719	接尾 s_22	{ion}_{N, V}_名詞_サ変名詞_*
-1.88757229578484398	接頭 s_21	大文字_ひらがな_名詞_人名

3.2 結果

結果を表 5 に示す。 3 つのシナリオに沿って新語を登録し、対訳素性を活性化しない単言語方式 (×) と活性化させた両言語方法（○）を試した結果を、表の左側に示す。精度は、全形態素、全カタカナ形態素、追加登録した新語で、調べた。ベースライン結果を、表の右側に示す。mzc– は、基本辞書のみを使って CRFs で解析した結果である。jmn– は、基本辞書のみを使って、JUMAN で解析した結果である。mozc– と jmn– では、新語は未知であり、その品詞は内蔵された未知語処理モジュールによって付与される。jmn+ は、理想的なシナリオで準備した新語辞書を追加して、JUMAN で解析した結果である。

いずれのシナリオにおいても、対訳素性を活性化しない単言語方式（×）の方が、結果が良い。原因は、2.3 節で述べた英訳と品詞の自動付与が、単純すぎたためではないかと考えている。英訳の探索は、文字列の完全一致で行なったため、同表記異語（グリーンピース, green peas, Green Peace）に対応できない。加えて、本来、訳語の品詞も曖昧であるが、これも無視した。表 6 に理想的なシナリオで学習された対訳素性のサンプルを載せる。品詞型テンプレートは singleton 素性も doubleton 素性も、おおよそ 0 付近の重さであり、解析に寄与していなかった。総合的に判断すると、単純な文字列完全一致による訳語と品詞の自動付与から展開した対訳素性は、あまり有効でないことがわかった。一つだけ例外がある。現実的なシナリオでは、両言語方式の方が、新語の結果が良い。追加する新語の支配的な品詞がわからない場合、単言語方式では全く判断材料がないが、対訳素性から手がかりを補っている、とも考えられる。カタカナ形態素及び新語の割合が少ないので、現段階では、はっきりと結論付けられない。

新語の品詞を普通名詞に固定した基本的なシナリオの方が、可能な品詞を列挙する現実的なシナリオより、結果が良い。現実的なシナリオでは、曖昧性を許容したため、ラティスの枝分かれが多くなり、学習困難に陥ったと考えられる。現状では、新語の支配的な品詞を一様に付与して追加する方が、得策であることがわかった。

4 まとめ

本稿では、借用語に適切な品詞を付与する問題を取り上げた。新たに対訳辞書から得られる素性を日本語形態素解析モデルに取り入れ、その有用性を調べた。見出し語の文字列完全一致を軸に対訳素性を展開し、形態素解析モデルに導入しても、効果がないことがわかった。

借用語の品詞付与は、品詞の大分類のみならず細分類までの一致を要求する厳しい問題である。しかしながら、本稿の試みは不完全で、改良の余地がある。第一に、英訳と品詞の自動付与の改良が挙げられる。例えば、訳語候補や品詞候補を同等に扱わず、尤もらしさで優先順位をつけるなどが考えられる。同時に、効果的な素性設計と素性選択が求められる。解析に使われた素性数 ($\lambda_k \neq 0$) は、単言語方式は 666,425 、両言語方式は 707,799 となり、肥大化する一方である。精度を向上するためにも高速化するためにも、効果的な素性設計と素性選択は、避けられない課題である。その他、いろいろあるが、すべて、今後の課題としたい。

謝辞 本研究は、工藤拓氏の実装を参考にしました。貴重な助言もいただきました。感謝いたします。言語資源を開発された方々にも、感謝の意を申し上げます。山田寛康氏、高村大也氏、乾孝司氏、坪井祐太氏から、有意義なコメントをいただきました。各氏に感謝申し上げます。

参考文献

[1] 内山 将夫, 井佐原 均 (2003). 日英新聞の記事および文を対応付けるための高信頼性尺度, 自然言語処理, vol.10-4, pp.201–220

[2] Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). Appliying Conditional Random Fields to Japanese Morphological Analysis, EMNLP, pp. 230–237

[3] 益岡隆志, 田窪行則 (1992). 基礎日本語文法 改定版, くろしお出版