Web コーパスを利用した辞書の派生訳語情報自動増殖

熊野 明　　梶 博行
静岡大学 創造科学技術大学院 自然科学系教育部 情報科学専攻

1. はじめに

ルールベースの機械翻訳では、精密な規則を記述することにより、多様な様態の訳文を生成することができる。品詞を転換したり、複数の語による表現を一つの語に写像したりする規則を用意することで、柔軟な訳文生成が可能になる。しかし、そのような規則に対応するためには、対訳辞書に品詞や表す意味の範囲が見出し語とは一致しない派生語も訳語として登録しなければならず、辞書の開発コストが増大する。大規模な商用システムでも辞書の網羅性は十分でないのが現状である。

本稿では、上記の問題を解決するため、既存の日英対訳辞書に含まれる派生訳語情報から派生変化規則を獲得する方法を提案する。獲得した規則を既登録の訳語に適用することにより新たに登録すべき派生訳語を生成する。精密な適用条件をもつ規則を獲得することは困難であり、派生訳語は過剰生成されるので、フィルタリングのために web コーパスを利用する。以下、提案方法について述べるとともに、市販の日英機械翻訳システムの対訳辞書と LDC Web コーパスを用いた実験について報告する。

2. 日英対訳辞書の訳語における派生関係

本研究では、さまざまな派生訳語を含む対訳辞書を対象とする。派生訳語には見出し語と品詞が異なる訳語や見出し語に表す意味に何らかの意味が加わった訳語が含まれる。これらの訳語は、機械翻訳システムが利用するために必要な情報、すなわち品詞および見出し語との意味のずれを表す素性とともに登録されている。図 1 に日英対訳辞書のエントリーの例を示す。ここで、[+able] は、可能の意味が付加されることを表す素性、[+not] は否定の意味が付加されることを表す素性である。

サ変動詞「適用する」に対しては、動詞訳語 apply のほか、名詞訳語 application、可能の意味が加わった形容詞訳語 applicable、同じく可能の意味が加わった名詞訳語 applicability、可能の意味と否定の意味が加わった形容詞訳語 inapplicable などが登録されている。

形容動詞「親切」に対しても、形容詞訳語だけでなく、副詞訳語、名詞訳語が登録されている。

機械翻訳システムはこの情報を利用して、日本語の動詞句を英語の名詞句で訳出したり、「適用する/こと/が/できる」という表現を can be applied に代えて applicable と訳したり、「適用/できる/か/どう/か/を」を if … can be applied に代えて applicability を使って訳したりすることができる[1]。

適用（サ変動詞）
apply (v)
application (n)
applicable (adj) [+able]
applicability (n) [+able]
inapplicable (adj) [+able][+not]

親切（形容動詞）
kind (adj)
unkind (adj) [+not]
kindly (adv) [+manner]
unkindly (adv) [+manner][+not]
kindness (n) [+state]
unkindness (n) [+state][+not]

図 1. 日英対訳辞書エントリーの例

3. 提案方法

3.1 概要

図 2 に示すように、提案方法は次の 3 つのステップから構成される。

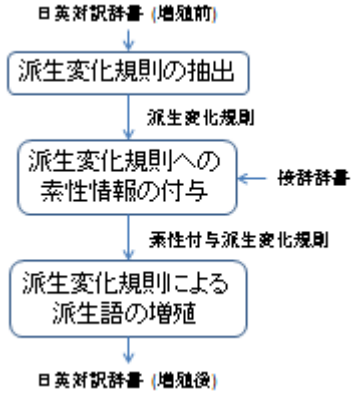

図 2. 提案方法の全体構成

(1) 派生変化規則の抽出

日英対訳辞書で同一見出しに含まれる英訳語の組は、派生関係にあると仮定し、接辞部分の違いから派生変化規則を抽出する。例えば、「適用」の訳語 apply(v) と application(n)の組から -ly (v) -> -lication (n) という規則が、apply(v)と applicable(adj)の組から -ly (v) -> -licable (adj) という規則がそれぞれ抽出される。

(2) 派生変化規則への素性情報の付与

接辞辞書と照合し、派生変化規則の言語的機能を明確にした素性付与派生変化規則にする。例えば、規則

-ly (v) -> -licable (adj) は接辞 able によって可能という意味が付与される規則であるとして、素性情報が付与された規則 -ly (v) -> -licable (adj) [+able] に高められる。

(3) 派生変化規則による派生語の増殖

日英対訳辞書の英訳語に素性付与派生関係規則を適用し、派生語を生成して追加する。例えば、「繁殖(サ変動詞)」の訳語として登録されている multiply(v)に規則 -ly (v) -> -lication (n) が適用され、multiplication(n) が追加登録される。「依存(サ変動詞)」の訳語 rely(v)と規則 –ly (v) -> -lication (n)から relication(v) が生成されるが、web コーパスを参照することによって実在しない語と判定される。

3.2 派生変化規則の抽出

同一見出し語に対する訳語には、相互に派生関係があると仮定する。例えば「回転(サ変動詞)」の訳語には、rotate(v), rotation(n)などがあり、「親切(形容動詞)」の訳語には kind(adj), kindly(adv), kindness(n)などがある。rotate(v)と rotation(n)は派生関係にあり、kind(adj), kindly(adv), kindness(n)も、相互に派生関係があると考える。ここで想定する相互の派生関係から、派生変化規則を抽出する。

3.2.1 訳語の組の文字アライメント

訳語相互に最小編集距離計算を求める DP アルゴリズムを用いて文字のアライメントを取り、後方の相違部分を取り出す[2]。

rotate と rotation の組に対して最小編集距離を得る例を示す。

		r	o	t	a	t	i	o	n
	0	1	2	3	4	5	6	7	8
r	1	0	1	2	3	4	5	6	7
o	2	1	0	1	2	3	4	5	6
t	3	2	1	0	1	2	3	4	5
a	4	3	2	1	0	1	2	3	4
t	5	4	3	2	1	0	1	2	3
e	6	5	4	3	2	1	1	2	**3**

白いセルが最小編集距離を与えるアライメントのパスを示す。相違部分は太枠で示した部分であり、最小編集距離は表右下隅に示す 3 である。共通部分と相違部分を以下に示す。

rotate = rotat(共通部分) + e(相違部分)
rotation = rotat(共通部分) + ion(相違部分)

また、派生関係にある英単語の間ではある程度の共通部分があるので、両者の編集距離と語長が長い単語長との比は 0.5 を上回らないと仮定する。「回転」の訳語である rotate と revolve の場合、編集距離が 5 であり、revolve の語長 7 に対する比が 0.71 なので、派生関係にあるとは認めない。

3.2.2 派生変化規則への変換

上述の文字アライメント処理を同一見出し語の訳語に属する任意の 2 つの訳語の組に行った結果を利用する。品詞の違いと相違部分を取り出すことで、以下のような対応関係が抽出できる。

〈語尾 1〉(〈品詞 1〉) -> 〈語尾 2〉(〈品詞 2〉)

これを、派生変化規則と呼ぶ。語尾が語尾 1 に一致する品詞 1 の英単語は、語尾が語尾 2 に置き替えることで、品詞 2 の単語に派生することを示す。

なお、派生変化規則の要素〈語尾 n〉は、相違部分だけでなく、共通部分の末尾の文字を 1 文字加えたものを使用することとした。

-te (v) -> -tion (n)
-d (adj) -> -dly (adv)
-d (adj) -> -dness (n)

例えば、-e (v) -> -ion (n) の派生変化は、rotate -> rotation, revise -> revision のように、直前の文字が s, t など一部の子音字に限られる。このように直前文字の条件を加えることにより、後の派生語生成処理で、arrange に対して arrangion のような明らかに不正な派生語候補の生成を避けることができる。

3.2.3 冗長な派生変化規則の除去

派生変化規則へ変換過程では kindly, kindness の組から

dly (adv) -> dness (n)

の規則が得られる。しかし、これは kind, kindly から得られる規則

-d (adj) -> -dly (adv)

と kind, kindness から得られる規則

-d (adj) -> -dness (n)

の規則に分割することができる。個々の派生変化規則は単独の意味を持つべきであるため、最初のものは派生変化規則から除去する。

なお、既存の日英対訳辞書には誤りを含む場合もあるので、機械的処理では誤った派生変化規則を抽出する可能性がある。そこで、抽出する規則の頻度に閾値を設け、それに満たないものは規則と認めないことにした。

また、接頭辞に相当する派生変化規則も抽出できるが、本稿では接尾辞だけを扱う。

3.3 派生変化規則への素性情報の付与

接辞の付与には規則性があり、接辞が付与することで接辞に対応する何らかの意味情報が加わる。MorphoQuantics は、英語の派生語を構成する接辞とその素性情報を、派生例とともに標準形で整理したデータである[3]。

このデータを利用して、派生変化規則に利用する接辞辞書を作成した。派生変化規則に対する素性情報の付与には、このデータを利用する。表 1 にそのデータの例を示す。接辞の標準形に対して、それが付加する対象語の品詞、付加して得られる派生語の品詞、接辞の付加で加わる意味素性の情報を収録している。

表 1. 接辞辞書のデータの例

接辞（標準形）	派生前品詞	派生後品詞	素性情報
-able	v, n	adj	+able
-ible	v, n	adj	+able
-ion	v	n	+action
-ly	adj	adv	+manner
-ness	adj	n	+state

3.2 で自動抽出した派生変化規則の語尾部分を、接辞の標準形と照合する。対応するものには、この情報を付与して素性付与派生変化規則と認める。対応しないものは規則とは認めず、以下の処理で使わない。少数の特殊な例から自動抽出された規則は派生関係を示すものではないので、ここで除去することができる。

接辞辞書に記述されている接辞は標準形であるため、実際の訳語から得られる派生変化辞書の語尾と一致しない場合がある。apply と applicable から得られる icable は接辞辞書の標準形としては存在しないが、標準形 able との編集距離 2 である。接辞辞書にある他の標準形との編集距離はこれより大きいので、最小編集距離をもつ able の変化形とみなすことができる。

照合して得られた接辞辞書のデータから、素性情報を派生変化規則にコピーして、素性付与派生変化規則とする。

派生変化規則の抽出と、素性付与処理の過程を図 3 に示す。

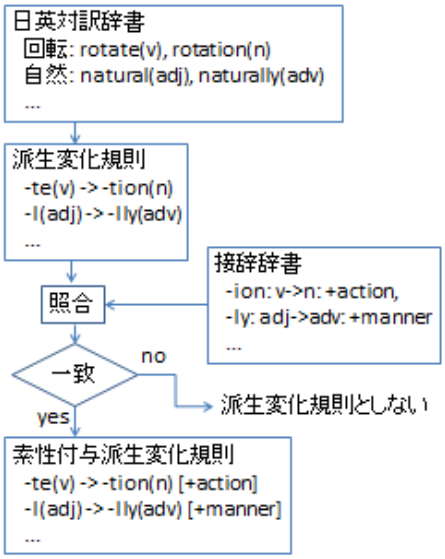

図 3. 素性付与派生変化規則の作成処理

3.4 派生変化規則による派生語の増殖

3.3 で獲得した素性付与派生変化規則を、日英対訳辞書の英訳語に対して適用し、派生語候補を生成する。適用条件は、変化前の語尾と品詞である。たとえば、cooperate(v)に対して、-te (v) -> -tion (n) の規則を適用することで、cooperation(n)が得られる。

この処理で生成された英単語が正しいものであれば、日英対訳辞書に追加する意味がある。正しい単語であるかどうかは、大規模な英語辞書で調べる必要があるが、web 検索することでも検証できる。本研究では実際の検索の代わりに web コーパスを利用する。

LDC Web コーパスデータは、web コーパス(延べ約 1 兆語)から、1-5-gram を頻度順に整理したデータであり、unigram データとして約 1,300 万語を含んでいる[4]。このデータと照合し、生成された派生語候補と一致するものがあれば実在する英単語とみなす。

図 4 は、この過程を示す図である。

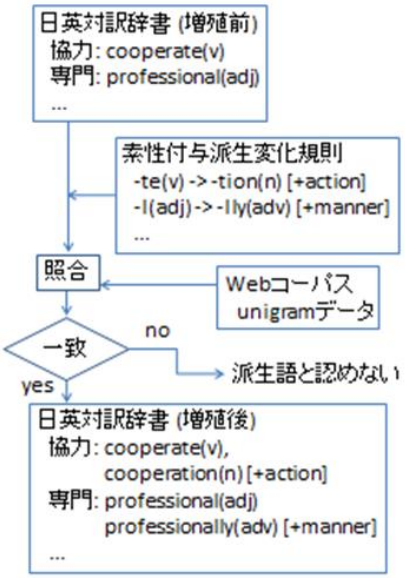

図 4. 派生語増殖の処理

英単語の派生には power -> powerful [+having] -> powerfully [+manner] のように、接尾辞が推移的に付加することがある。これを再現するため、派生語の増殖は 2 段階で行う。増殖前の日英対訳辞書の訳語に対して派生語生成を一度行い、正しいものを訳語として追加する。この追加した訳語に対して新たに派生語生成を行い、正しいものをさらに追加して増殖処理を行う。

4. 評価実験

提案方法を使って、日英対訳辞書の増殖実験を行い、派生語の増殖率とその精度を評価した。

4.1 実験データ

実験データには、商用の日英機械翻訳システムに使われている日英対訳辞書の一部を用いた。日本語見出し 9,229 語(サ変動詞 5,346 語、形容詞 738 語、形容動詞 3,145 語)に対する対訳データを増殖前の対訳辞書とした。他の品詞に比べ、英語での派生変化を多く含むものである。

日英対訳辞書から抽出する派生変化規則頻度の閾値は 2 とし、頻度 1 のものは規則としなかった。

4.2 実験結果

全 9,229 語の見出し語の訳語から抽出した素性付与派生変化規則の例を、頻度順に図 5 に示す。

1077 : -l (adj) -> -lly (adv) [+manner]
1058 : -e (adj) -> -ely (adv) [+manner]
965 : -te (v) -> -tion (n) [+action]
965 : -tion (n) -> -te (v) [-action]
834 : -t (adj) -> -tly (adv) [+manner]
700 : -e (adj) -> -eness (n) [+state]
644 : -s (adj) -> -sly (adv) [+manner]
500 : -d (adj) -> -dly (adv) [+manner]
458 : -s (adj) -> -sness (n) [+state]
410 : -ly (adv) -> -le (adj) [-manner]
410 : -le (adj) -> -ly (adv) [+manner]
396 : -nce (n) -> -nt (adj) [-action]
396 : -nt (adj) -> -nce (n) [+action]

図 5. 素性付与派生変化規則の例

最初の抽出処理で 7,696 種類の派生変化規則が得られた。うち、頻度 2 以上のものは 3,511 種類である。この中には「回復(サ変動詞)」の訳語 unrevivable, unrestorable 等から得られた -evivable (adj) -> -estorable (adj) のような規則も含まれていた。これらは、接辞辞書との照合で除去できた。

表 2 は、3.4 で示した派生語増殖処理を、日英対訳辞書の見出し語「哀願(サ変動詞)」の訳語に適用した結果である。第 1 段階の増殖結果と、その中から正しいものを選んだ後の第 2 段階の増殖結果を示す。

増殖前の訳語は、entreaty(n), appeal(n), entreat(v), implore(v)の 4 語である。まず第 1 段階の増殖では、訳語 appeal(n) から appealing(n), appealed(adj), appealable(adj)の 3 語が生成された。entreaty(n), entreat(v), implore(v)の 3 語からもそれぞれ派生語候補が生成された。appeal(n)からの派生語の 1 つである appealing(n)からは、第 2 段階で appealingly(adv), appealling(n)、appeale(v)の 3 語が生成された。

表 2. 派生変化規則を使った派生語増殖の例

増殖前	増殖語(第 1 段階)	増殖語(第 2 段階)
entreaty(n)	entreaties(n)	-
	entreate(adj)	-
appeal(n)	appealing(n)	appealingly(adv)
		appealling(n)
		appeale(v)
	appealed(adj)	appeals(n)
		appelation(n),
		appeales(n)
	appealable(adj)	appealability(n)
entreat(v)	entreated(adj)	-
	entreating(n)	-
	entreats(n)	-
	entreatment(n)	entreatments(n)
implore(v)	implored(adj)	-
	imploring(n)	implorings(n)
		imploringly(adv)
	implores(n)	-
	imploration(n)	*implorar(adj)*
		implorations(n)
	implorer(n)	-

(斜体のスペルは、誤った派生語を示す)

増殖前に entreaty(n), appeal(n), entreat(v), implore(v)の 4 種類だった訳語に対して、2 段階の増殖処理を通して派生語候補 26 語が生成された。このうちの 5 語、entreate(adj), appealling(n), appeale(v), appeales(n), implorar(adj)は英単語として正しくないため、正しい派生語はこれらを除く 21 語である。

サ変動詞 10 語の訳語増殖処理に対する同様の評価を示す。

増殖前訳語総数	52 語
生成された派生語候補	571 語
正しい派生語	366 語

増殖前の訳語 52 語に対して 7 倍の派生語を追加することができた。派生増殖の少ない例では 3 語に対して 9 語、多い例では 3 語に対して 45 語増殖した。

4.3 結果の検討

多様な訳語を増殖できる可能性を示すことができたが、その精度にはまだ課題が残る。

正しくない訳語の多くは、web コーパスの unigram の精度によるものである。web コーパスは実際の web ページから収集したものであるために、少数のミススペルは避けられない。この結果、誤った英訳語を実在する語と判断してしまう。これを避けるためには、低頻度語を検証に使用しない方法がある。今後、web コーパスとの照合方法を検討する。

その他にも、派生変化規則の適用を誤ったものもあった。たとえば、expect, expected の組から得られた規則 -t (v) -> -ted (adj) を適用し、sit(v)から sited(adj)を派生語として生成した。sited は site(v)の過去形であり英単語として誤りではない。しかし sit の派生語としては正しくない。これを避けるためには、派生変化規則の適用条件の改良が必要である。

5. まとめと今後の課題

既存の日英対訳辞書の派生訳語情報を利用して、機械的に派生変化規則を獲得した。既登録の英訳語に適用することで、派生訳語情報を自動的に増殖できる見通しが得られた。この方法で得られる派生語は、多様な品詞や意味範囲の語を含んでいる。日英機械翻訳において多様な様態の訳文を生成するために活用することができる。

しかし、機械的に派生変化した語の正当性に対しては課題が残る。最終的には人手のチェックが欠かせないが、意味付与派生変化規則抽出、派生語候補生成の精度を向上させることで、日英対訳辞書の派生語情報増殖の効率を高める方法を検討する。

【参考文献】

[1] Kaji, H; Dictionary Structure for Flexible Lexical Transfer in Machine Translation, International Symposium on Electronic Dictionaries (1988)
[2] Gusfield, Dan; Algorithms on strings, trees, and sequences. Cambridge University (1997
[3] Laws, J.V. & C. Ryder: MorphoQuantics: <http://morphoquantics.co.uk> (2014)
[4] LDC: Web 1T 5-gram Version 1: <https://catalog.ldc.upenn.edu/LDC2006T13> (2006)