Wikipedia と要素合成法・編集距離を用いた専門用語の訳語推定

釜田郁也　綱川隆司　梶博行
静岡大学大学院情報学研究科
gs13011@s.inf.shizuoka.ac.jp, {tuna, kaji}@inf.shizuoka.ac.jp

1. はじめに

機械翻訳であれ人間翻訳であれ、専門分野の文書の翻訳においては、専門用語を正しく翻訳することが重要である。しかし、専門用語の多くは一般の対訳辞書に登録されていない。このため、専門用語の訳語を知るために Web を活用することが考えられる。本稿では Wikipedia を利用した専門用語の訳語推定方法を提案する。

Wikipedia を利用する理由は以下のとおりである。Wikipedia は各分野の専門家の共同作業によって編集されており、少なくとも専門用語については Web 全般より信頼度が高いと思われる。絶えず記事内容が更新され、記事数も増加しているので、最新の専門用語が含まれる可能性が高いという利点もある。また、同じような内容を記述した各言語版の記事がリンクで結ばれた、いわゆるコンパラブルコーパスであるので、対訳抽出に適した言語資源である。

専門用語は複合語が多く、対訳語の間で構成要素の対応関係が成立することが多い。このため、専門用語の訳語推定には要素合成法がよく用いられる(藤井，石川，2000; Baldwin and Tanaka, 2004)。しかし、それに必要な構成要素対訳辞書のカバー率が問題となる。この問題を解決するため、本稿では、合成される訳語だけでなくそれと類似の文字列も訳語候補と考え、原言語の専門用語を含む記事とリンクで結ばれた目標言語の記事を探索する方法を提案する。提案方法は基本的に言語対に依存しないが、本稿では原言語を日本語、目標言語を英語として記述する。

2. 関連研究

要素合成法における構成要素対訳辞書のカバー率の問題を解決するため、外池ら(2007)は対訳複合語のサンプルから構成要素の対応を獲得、利用する方法を提案した。例えば、(応用数学, applied mathematics)、(応用科学, applied science)といったサンプルから、一般の対訳辞書には登録されていない対応関係(応用，applied)を獲得する。小松原ら(2012)は、コンパラブルな文書対の集合から相関値付き対訳辞書を獲得し（文書対中の共起頻度に基づく相関値を計算することにより(応用，applied)のような対応関係が得られる）、これを参照して複合語に対し相関値付きの訳語候補を生成する方法を提案した。これら 2 つの方法の効果は、利用できる対訳複合語のサンプルあるいはコンパラブルな文書対の量に依存する。

Tsuruoka and Tsujii (2003)は、生物医学分野の文書からの情報抽出の研究の中で、編集距離を用いてコーパスから異表記の語を検索する方法を提案している。

3. 着眼点

EDR 辞書を含む多くの対訳辞書において、名詞“応用”の訳語として形容詞“applied”（または動詞“apply”）は登録されていない。このため、要素合成法による複合語“応用行動分析”の訳語は“application behavior analysis”などであり、正しい訳語“applied behavior analysis”は含まれない。しかし、“application behavior analysis”と“applied behavior analysis”は類似の文字列である。“application”と“applied”は派生語の関係にあり、接辞の置換、削除、挿入といった若干の編集によって一方から他方に変換することができる。この例のように、要素合成法による合成訳が正しい訳語でなくても、それとの編集距離が一定の範囲に正しい訳語が存在する可能性は高い。

要素合成法による訳語推定では、合成訳が正しい訳語であることを検証するためにコーパスが用いられるが、Wikipedia はこの目的に好都合である。訳語を推定すべき専門用語を含む原言語の記事とリンクで結ばれた目標言語の記事に当該専門用語の訳語が含まれている可能性が高いからである。訳語を推定すべき専門用語を含む原言語の記事を特定し、それらに対応した目標言語の記事のなかで、要素合成法による訳語候補と編集距離が小さい単語列を探索すればよい。

4. 提案方法

提案方法は、(1)要素合成法による訳語候補の生成、(2)訳語の検証に用いる記事の選択、(3)訳語候補と類似の単

語列の探索の 3 ステップから構成される。以下、各ステップの詳細を記述する。

4.1 要素合成法による訳語候補の生成

入力タームxを単語に分割する。xがn個の単語の列$w_1 w_2 \ldots w_n$であるとき、以下の手続きにより、三角行列$\mathrm{A}(i,j)(1 \le i \le n, i \le j \le n)$を作成する。図 1 に例を示す。

1) w_iの訳語が対訳辞書に登録されているなら、それらを対角線上のセル$\mathrm{A}(i,i)$に書き込む（$i = 1,2,\cdots,n$）。
2) 対角線近くから遠くの順にセル$\mathrm{A}(i,j)$を選び、
   a) 複合語$w_i w_{i+1} \ldots w_j$の訳語が対訳辞書に登録されているなら、それらをセル$\mathrm{A}(i,j)$に書き込む。
   b) ２つの単語列$y\big(\in \mathrm{A}(i,k)\big)$と$y'\big(\in \mathrm{A}(k+1,j)\big)$をこの順に並べて得られる単語列を全て$\mathrm{A}(i,j)$に追加する（$k = i,\cdots,j-1$）。

この手続きにより$\mathrm{A}(1,n)$に得られる単語列が入力タームに対する訳語候補である。

上の手続き中の 2b)は、複合語内の構成要素の順序が言語間で一致することを前提としている。日本語と英語の間では構成要素の順序が一致する複合語が大部分である。しかし、言語対によって、構成要素の順序が言語間で一致するとは限らない。例えば、日本語と仏語の間では(自動翻訳, traduction automatique)のように順序が逆になることが多い。そのような言語対の場合、yとy'の順序を逆にした単語列も追加することとする。

構成要素として望ましい単位は必ずしも一致しない。例えば、日本語では接辞（例えば、“一般的”の“的”、“一般化”の“化”）を独立の構成要素と考えることによって辞書のエントリを少なくすることができる。しかし、英語でそれらに対応する要素を分離するのは必ずしも容易でない（例えば、“一般的”に対応する“general”では“的”の意味を“形容詞”という品詞が担っている。“一般化”に対応する“generalization”では“化”の意味を接尾辞“ize”（接辞“tion”との接続のためさらに“iza”に変化）が担っている）。このため、原言語の特定の構成要素（例えば、日本語の“的”、“化”）の訳語として、訳出しないことを意味する nil を加えることにする。

<table>
<tr><td rowspan="3">応用</td><td rowspan="2">・application</td><td>・application behavior
・application action</td><td>・applicaton behavior analysis
・application action analysis</td></tr>
<tr><td rowspan="2">・behavior
・action</td><td rowspan="2">・behavior analysis
・action analysis</td></tr>
<tr><td>行動</td></tr>
<tr><td colspan="2"></td><td>分析</td><td>・analysis</td></tr>
</table>

図 1　訳語候補の生成例

4.2 訳語の検証に用いる記事の選択

入力言語の Wikipedia 記事のうち、入力タームが出現する記事を選択する。選択された記事の各々について、目標言語の記事への言語間リンクをもつかどうかチェックする。目標言語の記事への言語間リンクをもつなら、リンク先の記事を訳語の検証に用いる記事として選択する。

4.3 訳語候補と類似の単語列の探索

4.2 で得られた目標言語の記事集合から単語 N グラムを抽出し、4.1 で生成された訳語候補$y, y_2, \cdots, y_k$との類似度を計算する。訳語候補y_iと単語 N グラムyの類似度を次式で定義する。

$$\mathrm{Sim}(y_i, y) = 1 - \frac{\mathrm{d}(y_i, y)}{\max\{|y_i|, |y|\}}$$

ここに、$\mathrm{d}(y_i, y)$はy_iとyのレーベンシュタイン編集距離 (Jurafsky and Martin, 2009)、$|y_i|$、$|y|$はそれぞれy_i、yの長さ（文字数）である。

2 つの文字列の編集距離は、一方を他方に変換するのに必要な編集操作（文字の置換、削除、挿入）の数の最小値である。2 つの文字列$a_1 a_2 \ldots a_m$と$b_1 b_2 \ldots b_n$の編集距離はダイナミックプログラミングの手法を用いて次のように計算される。C は、$(m+1)$行、$(n+1)$列の行列で、$\mathrm{C}(i,j)$は部分文字列$a_0 a_1 \ldots a_i$と$b_0 b_1 \ldots b_j$（a_0とb_0は空白）の編集距離を表す。図 2 に“application behavior analysis”と“applied

		a	p	p	l	i	e	d		b	・・・	i	s
	0	1	2	3	4	5	6	7	8	9	・・・	24	25
a	1	0	1	2	3	4	5	6	7	8	・・・	23	24
p	2	1	0	1	2	3	4	5	6	7	・・・	22	23
p	3	2	1	0	1	2	3	4	5	6	・・・	21	22
l	4	3	2	1	0	1	2	3	4	5	・・・	20	21
i	5	4	3	2	1	0	1	2	3	4	・・・	19	20
c	6	5	4	3	2	1	1	2	3	4	・・・	19	20
a	7	6	5	4	3	2	2	2	3	4	・・・	18	19
t	8	7	6	5	4	3	3	3	3	4	・・・	18	19
i	9	8	7	6	5	4	4	4	4	4	・・・	17	18
o	10	9	8	7	6	5	5	5	5	5	・・・	16	17
n	11	10	9	8	7	6	6	6	6	6	・・・	15	16
	12	11	10	9	8	7	7	7	6	7	・・・	15	16
b	13	12	11	10	9	8	8	8	7	6	・・・	15	16
・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・・・	15	16
i	28	27	26	25	24	23	22	22	21	21	・・・	6	7
s	29	28	27	26	25	24	23	23	22	22	・・・	7	6

図 2　編集距離の計算例

behavior analysis”の編集距離の計算を例示する。

1) $C(i,0) \leftarrow i \ \ (i = 0,1,\cdots,m)$

2) $C(0,j) \leftarrow j \ \ (j = 0,1,\cdots,n)$

3) （行列の上から下、左から右の順に）

$a_i = b_j$なら、 $C(i,j) \leftarrow C(i-1,j-1)$

$a_i \neq b_j$なら、 $C(i,j) \leftarrow$

$\min\{C(i-1,j), C(i,j-1), C(i-1,j-1)\}+1$

ここで、訳語推定精度を高めるとともに処理時間を短縮するため、類似度を計算する訳語候補y_iと単語 N グラムyに次のような制限を加える。

- 記事集合から抽出される単語 N グラムの中には語として不適切な品詞の並びも多いので、明らかに不適切な N グラム（具体的には、先頭の単語が前置詞あるいは冠詞の N グラム、末尾の単語が前置詞、冠詞あるいは動詞の N グラム）を候補から除外する。
- 訳語候補y_iと単語 N グラムyの単語数がかけ離れている場合、$\mathrm{Sim}(y_i, y)$が大きいことは期待できないので、単語数の差が 2 以下の組合せについてのみ類似度を計算する。

上記の制限の下で、訳語候補y_iと単語 N グラムyの類似度を計算する。$\max_{y_i} \mathrm{Sim}(y_i, y)$を単語 N グラム$y$のスコアと考え、このスコアの降順リストを出力する。表 1 に出力例を示す。入力ターム“応用行動分析”に対して要素合成法で生成される訳語候補“application behavior analysis”を経由して正しい訳語“applied behavior analysis”が第 1 位にランクされている。

表 1　類似単語列の探索結果の例

入力ターム	訳語候補	Nグラム	スコア	正誤
応用行動分析	application behavior analysis	applied behavior analysis	0.793	○
		applied behavior analyst	0.724	×
		practice of behavior analysis	0.689	×

5. 評価実験

5.1 使用データ

2014 年 5 月時点の Wikipedia のうち、言語間リンクで結ばれた日本語と英語の記事対、346717 組を使用した。要素合成法に用いる日英対訳辞書は、EDR 対訳辞書、EDICT、英辞郎の日英対訳対、さらに言語間リンクで結ばれた Wikipedia 記事対のタイトルの組をマージして作成した。

テストセットとして、デジタル人工知能学事典（人工知能学会編, 共立出版, 2008）と言語処理学事典（言語処理学会編, 共立出版, 2010）の和英索引からランダムに 2519 対を選択した。それらの日本語タームを入力タームとし、評価の際、英語タームを正解訳語と考えた。

2519 対のテストセットのうち、日本語タームが Wikipedia に含まれているものが 1382 対、英語タームが Wikipedia に含まれているものが 1756 対、日本語タームと英語タームがともに Wikipedia に含まれているものが 1281 対、日本語タームと英語タームを含む Wikipedia 記事対が存在するものが 969 対であった。

5.2 実験結果

次の 3 案の提案方法と従来方法を実行し、正解訳語の平均逆順位 MRR (Mean Reciprocal Rank)を算出した結果を表 2 に示す。

(a) 編集距離に基づく類似単語列の探索のみで、接尾辞に対する nil 訳語と N グラムの品詞列制限は採用しない。

(b) (a)＋接尾辞に対する nil 訳語

(c) (b)＋N グラムの品詞列制限

なお、提案方法と従来方法を比較することが目的であるので、日本語タームと英語タームを含む Wikipedia 記事対が存在する 969 対を対象として MRR を計算した。対訳知識源としての Wikipedia の有用性を切り離した評価であることに注意されたい。

従来の要素合成法は、対訳辞書を参照して得られる合成訳のうち、入力タームが出現する日本語記事と結ばれた英

表 2　実験結果

方法		k位以内に正解が得られたターム数						MRR
		k=1	2	3	4	5	∞	
提案方法	(a)編集距離	429	493	525	544	556	578	0.494
	(b)編集距離+接尾辞nil訳語	477	551	574	585	593	662	0.545
	(c)編集距離+接尾辞nil訳語+Nグラム品詞列制限	484	555	582	597	609	660	0.554
従来方法（編集距離なし）		373.8	414.6	430.9	440.5	446.1	452	0.416

語記事中に出現するものを出力する。複数の訳語が出力されることもあるが、訳語に順位はつかない。そこで、MRR は次のように算出した。k個の訳語が出力されたとき、どの訳語も第 1 位、第 2 位、…、第k位にそれぞれ確率$1/k$でランクされると考えられる。よって、正解が含まれるとき、$\mathrm{MRR} = (1/k)(1 + 1/2 + \cdots + 1/k)$。正解が含まれないとき、$\mathrm{MRR} = 0$。

ここで、テストセットの正解データについてコメントしておく。人工知能学事典、言語処理学事典の和英索引に記載された訳語以外は誤りと判定したが、実際は正しい訳語であるというケースがあった。例えば、“エピソード記憶”に対する訳語リストの上位にランクされた“episodic memory”は正解訳語“episode memory”と異なるので誤りと判定されたが、この判定は正しくない。このような例があるため、表 2 の MRR は過小評価となっている可能性がある。

5.3 結果の検討

表 2 から、編集距離に基づく類似単語列の探索を加えることによって要素合成法の能力が向上したことがわかる。接尾辞に対して nil 訳語を認めること、目標言語の N グラムの品詞列を制限することの効果も確認できた。しかし、正しい訳語が第 1 位にランクされた入力タームはほぼ半分であり、まだまだ改良が必要である。

MRR の値を低下させる最も大きな要因となったのは、要素合成法による訳語候補生成ステップの失敗である。969 の入力タームのうち、“カルマンフィルタ”、“最汎単一化子”（下線を引いた要素が対訳辞書に未登録）など 165 語では、訳語候補が得られなかったため、類似単語列の探索もできず訳語リストが出力されなかった。また、要素合成法によって生成される訳語候補が正解訳語とかけ離れているため、類似単語列探索の効果が得られない例もあった。要素合成法が参照した対訳辞書に含まれる“遺伝”の訳語は“inheritance”、“heredity”だけで、“gene”から派生した語は含まれていなかった。このため、入力ターム“遺伝的要因”に対して正解訳語“genetic factor”が上位にランクされることにはならなかった。以上の例からも、要素合成法において参照する対訳辞書のカバー率が依然として大きな課題であるといえる。片仮名語に対する翻字も必要と思われる。

5.1 に示したように、テストセットに対する Wikipedia のカバー率はあまり高くないが、このことが提案方法の有効性を否定することにはならない。Wikipedia は拡大し続けているからである。入力タームと正解訳語を含む記事対が得られる比率は徐々に高くなるであろう。なお、今回のテストセットには、そもそも Wikipedia 記事中にタームとして出現することが期待できない日本語タームが含まれていた。例えば、(有限資源下のプランニング, resource bounded planning)がその例で、英語タームを直訳したものと思われる。このような例も考慮すると、訳語の検証に用いる記事を“入力タームの構成要素を含む原言語記事とリンクで結ばれた目標原言語記事”に変更したほうがよいかもしれない。

6. おわりに

編集距離を利用した類似単語列の探索機能で強化された要素合成法によって Wikipedia から専門用語の訳語を推定する方法を提案した。人工知能と言語処理分野のタームを対象とした評価実験では、0.554 の MRR を達成し、類似単語列の探索を行わない従来の要素合成法の 0.416 を上回った。

類似単語列探索の有効性は明らかになったが、要素合成法による訳語候補生成の重要性は変わらない。正しい訳語でなくてもそれに近い訳語候補を生成することが重要であるので、対訳辞書のカバー率向上が依然として課題である。また、検証に用いる Wikipedia 記事対の選択方法にも改良の余地がある。

参考文献

Baldwin, Timothy and Takaaki Tanaka. 2004. Translation by machine of compound nominals: Getting it right. In *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, pp. 24-31.

Jurafsky, Daniel and James H.Martin. 2009. *Speech and laguage processing, 2nd edition*. p.108, Prentice Hall.

Tsuruoka, Yoshimasa and Jun'ichi Tsujii. 2003. Boosting precision and recall of dictionary-based protein name recognition. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pp.41-48.

小松原慶啓, 綱川隆司, 梶博行. 2012. コンパラブルコーパスと Web を用いた用語翻訳器. 言語処理学会第 18 回年次大会発表論文集, pp.685-688.

外池昌嗣, 宇津呂武仁, 佐藤理史. 2007. ウェブから収集した専門分野コーパスと要素合成法を用いた専門用語訳語推定. 自然言語処理, Vol.14, No.2, pp.33–68.

藤井敦, 石川徹也. 2000. 技術文書を対象とした言語横断情報検索のための複合語翻訳. 情報処理学会論文誌, Vol.41, No.4, pp.1038-1045.