日英・中英対訳辞書からの日中対訳辞書の構築

綱川 隆司† 岡崎 直観† 辻井 潤一†‡§
† 東京大学大学院情報理工学系研究科コンピュータ科学専攻
‡ School of Computer Science, University of Manchester
§ National Centre for Text Mining, UK
{tuna,okazaki,tsujii} at is.s.u-tokyo.ac.jp

1 はじめに

本稿では，日英対訳辞書と中英対訳辞書から英語を介して日本語・中国語間の対訳辞書を自動的に構成するための統計的手法を提案する．対訳辞書は複数の言語を扱う際には不可欠の言語資源であるが，実際に存在する対訳辞書は英語とそれ以外の言語の組み合わせのものが大半であり，英語以外の二言語間においては対訳辞書が存在しないか，あるいは規模が小さい．また，専門用語の対訳辞書に関しては英語に対してのみ提供されている場合が多い．そこで，本研究では，英語を介して英語以外の言語間の対訳辞書を構成し，既存の辞書よりカバー率の高い辞書を構築する．

言語 L_e, L_f 間に対訳辞書を構築する際に，ピボットとなる第三言語 L_p を利用する手法は，これまでにも提案されてきた [9, 1, 8, 6, 7, 12, 2]．この手法は L_e, L_f 間に直接の資源がなくても適用可能という利点があるが，語の「曖昧性」および「不一致」という二つの主な問題点を解決する必要がある．

一般的に，言語 L_f の語 w_f と言語 L_p の語 w_p，および w_p と言語 L_e の語 w_e がそれぞれ対訳関係にあるとき，w_f と w_e が対訳関係にあるとは限らない．この問題は w_p に複数の意味がある際に生じやすい．例えば，日本語の「土手」と中国語の「银行」（銀行）は，ピボットとなる語 “bank” によって関係づけられてしまう．田中ら [13] は，訳語関係のグラフや語の形態素が持つ表意を利用して正しい訳語を選択する手法を提案している．Bond ら [1] はシソーラスを用いて訳語選択の際に意味クラスが近いものを選んでいる．Shirai ら [8] はピボットとして用いることのできる語数に関する調査を行っている．また，Schafer ら [7] は文脈の類似度，編集距離，語の相対頻度および burstiness 類似度を用いている．

一方，もう一つの問題は，独立に開発された 2 つの辞書をピボット言語 L_p で統合する際，同一のエンティティに対して異なる表現が用いられているため，辞書の統合がうまくいかないことである．例えば，日英辞書に「地球温暖化: global warming」，中英辞書に「全球变暖: global heating」と記述されている場合，これらを関係づけることができない．また，専門用語に関する対訳辞書においては，対応する訳語が存在しない場合が多い．

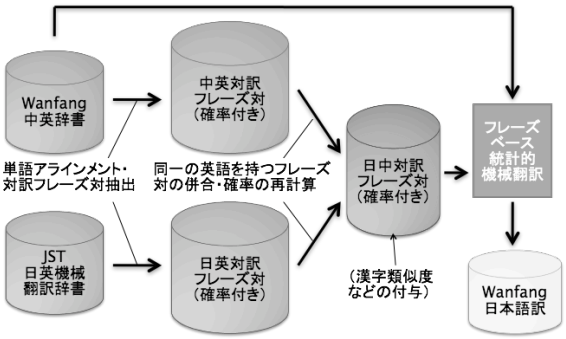

図 1: 日中対訳辞書構成のフレームワーク

本稿では，日中対訳辞書を構成するために，英語をピボット言語とし，フレーズベース統計的機械翻訳 [10, 11] を適用して，語の不一致の問題を改善する手法を提案する．統計的機械翻訳を適用するためには訓練データとして並行コーパスが必要であるが，本手法では日英辞書と中英辞書を並行コーパスとみなし，ここから抽出した日英および中英間の対訳フレーズ対を英語を介して日中対訳フレーズ対とすることで，機械翻訳に必要なフレーズ対を生成する．（図 1 を参照）．また，対数線形モデルを採用し，対訳関係を導くための有効な手掛かりである漢字の類似性 [12] を用いた．以下では，まずピボット言語を用いた対訳辞書の構成，およびその手法としての統計的機械翻訳について述べる．次に，日英および中英の専門用語および固有名詞を含む対訳辞書を用い，日中対訳辞書を構成する実験を行う．

2　2つの対訳辞書の結合

L_e, L_p, L_f をそれぞれ翻訳元言語，ピボット言語，翻訳先言語の語彙とし，以下の2つの対訳辞書 L_f–L_p および L_p–L_e があると仮定する:

$$L_f\text{–}L_p = \{(\bar{w}_f, \bar{w}_p) | \bar{w}_f \text{ は } \bar{w}_p \text{ の翻訳 }\}, \quad (1)$$

$$L_p\text{–}L_e = \{(\bar{w}_p, \bar{w}_e) | \bar{w}_p \text{ は } \bar{w}_e \text{ の翻訳 }\}, \quad (2)$$

ただし，$\bar{w}_x$ は語彙 L_x に含まれる単語の列とする．

2.1　単純な辞書引きによる結合

2つの対訳辞書 L_f–L_p および L_p–L_e から対訳辞書 L_f–L_e を構築する最も単純な方法は，共通の訳語 $\bar{w}_p$ を持つ2つの語 $\bar{w}_f$ と $\bar{w}_e$ を結びつけることである:

$$\begin{aligned} L_f\text{–}{L_e}^{(e)} &= \{(\bar{w}_f, \bar{w}_e) | \exists \bar{w}_p \ ((\bar{w}_f, \bar{w}_p) \in L_f\text{–}L_p \\ &\quad \wedge (\bar{w}_p, \bar{w}_e) \in L_p\text{–}L_e)\}. \end{aligned} \quad (3)$$

これを「単純な辞書引き」による辞書の結合と呼ぶことにする．

2.2　フレーズベース統計的機械翻訳を用いた結合

本研究では，ピボット言語を用いたフレーズベース統計的機械翻訳 [10, 11] を用いて2つの対訳辞書の結合を行う．まず，対訳辞書 L_f–L_p および L_p–L_e をそれぞれ並行コーパスとみなし，GIZA++および両方向のアラインメントを結合するための refinement method [5] を適用し，単語アラインメントを得る．そして，単語アラインメントに矛盾しない対訳フレーズ[1]対を抽出し，その相対頻度を確率値として保持する:

$$p(\bar{w}_f | \bar{w}_p) = \frac{C(\bar{w}_f, \bar{w}_p)}{C(\bar{w}_p)} \quad (4)$$

さらに，語彙 L_f と L_e に含まれる単語列の全ての組合せについて，以下の確率を計算することで，L_f と L_e の単語列間の翻訳確率を求める:

$$p(\bar{w}_f | \bar{w}_e) = \frac{\sum_{\bar{w}_p} p(\bar{w}_f | \bar{w}_p) p(\bar{w}_p | \bar{w}_e)}{\sum_{\bar{w'}_f} \sum_{\bar{w}_p} p(\bar{w'}_f | \bar{w}_p) p(\bar{w}_p | \bar{w}_e)}, \quad (5)$$

$$p(\bar{w}_e | \bar{w}_f) = \frac{\sum_{\bar{w}_p} p(\bar{w}_e | \bar{w}_p) p(\bar{w}_p | \bar{w}_f)}{\sum_{\bar{w'}_e} \sum_{\bar{w}_p} p(\bar{w'}_e | \bar{w}_p) p(\bar{w}_p | \bar{w}_f)}. \quad (6)$$

対数線形モデルを用いたフレーズベース統計的機械翻訳では，以下の式に基づいて，翻訳元の文 $\bar{w}_f$ を翻訳先の文 $\hat{\bar{w}}_e$ に翻訳する．

$$\begin{aligned} \hat{\bar{w}}_e &= \operatorname*{argmax}_{\bar{w}_e} \Pr(\bar{w}_e | \bar{w}_f) \\ &= \operatorname*{argmax}_{\bar{w}_e} \sum_{m=1}^{M} \lambda_m h_m(\bar{w}_e, \bar{w}_f), \end{aligned} \quad (7)$$

ただし，$h_m(\bar{w}_e, \bar{w}_f)$ は $\bar{w}_e, \bar{w}_f$ 間の対訳らしさを表す素性関数であり，λ_m はそれらの重みである．本研究では，統計的機械翻訳システムで文を翻訳するのではなく，語彙 L_f に含まれる単語列 $\bar{w}_f$ を翻訳することで，対訳辞書 L_f–L_e を導く．

式7の素性関数として，以下のものを用いた．

1. フレーズ翻訳確率
   $h_1(\bar{w}_e, \bar{w}_f) = \sum_i \log p(\bar{w}_e^{(i)} | \bar{w}_f^{(i)})$.
2. 翻訳先言語の 3-gram 言語モデル
   $h_2(\bar{w}_e, \bar{w}_f) = \log p(\bar{w}_e)$.
3. フレーズの並び替えスコア
   $h_3(\bar{w}_e, \bar{w}_f) = \sum_i d(\bar{w}_e^{(i)}, \bar{w}_f^{(i)})$.
4. 漢字の類似度
   $h_4(\bar{w}_e, \bar{w}_f) = \sum_i \mathrm{ksim}(\bar{w}_e^{(i)}, \bar{w}_f^{(i)})$.

ただし，$\bar{w}_e^{(i)}$ および $\bar{w}_f^{(i)}$ は，翻訳の際に用いた i 番目のフレーズ対を表す．

2.2.1　フレーズの並び替えスコア

先行研究 [4] に基づき，翻訳の際に用いたフレーズ対の順序が入れ替わった場合のペナルティを，次式で定義する．

$$d(\bar{w}_e^{(i)}, \bar{w}_f^{(i)}) = |a_i - b_{i-1} - 1|. \quad (8)$$

ここで，a_i は $\bar{w}_f^{(i)}$ の最初の単語の位置，b_{i-1} は $\bar{w}_e^{(i-1)}$ の最後の単語の位置（$i = 0$ のときは0）とする．

2.2.2　漢字の類似度

中国語と日本語はともに漢字を用いており，対訳関係の手掛かりとして漢字の共通性を用いることができる [12]．本研究では漢字の共通性を素性関数として導入するために以下の式を用いた:

$$\mathrm{ksim}(\bar{w}_e^{(i)}, \bar{w}_f^{(i)}) = \frac{\bar{w}_e^{(i)} \text{と} \bar{w}_f^{(i)} \text{の漢字の一致数}}{\bar{w}_e^{(i)} \text{と} \bar{w}_f^{(i)} \text{の漢字数の最小値}}. \quad (9)$$

ただし，漢字の一致については，簡体字と日本語の漢字の間で字形のみが異なるもの（例えば「汉」と「漢」）は同一とみなした．

[1] ここでのフレーズとは，構文的意味を持つ句ではなく，単語列を表すとする．

表 1: 中英・日英対訳辞書の語彙数と単純な辞書引きによって得られる対訳数

対訳辞書	L_C 語彙数	L_E 語彙数	L_J 語彙数
L_C–L_E	375,990	429,807	-
L_E–L_J	-	418,044	465,563
L_E 総語彙数	-	783,414	-
L_C–$L_J{}^{(e)}$	98,537 (22.4%)	68,996	103,437 (22.2%)

3 実験

本手法の有効性を検証するために，専門用語を含む中英および日英対訳辞書を用い，辞書統合実験を行った．用いた対訳辞書は以下の通りである．

中英 L_C–L_E 万方数据 (Wanfang Data)[2]英汉-汉英科技大词库[3]: 525,259 項目

日英 L_E–L_J JST 機械翻訳辞書[4]: 527,206 項目

表 1 はこれらの辞書が含む語彙数と，単純な辞書引きによって得られる中日辞書の語彙数を示している．単純な辞書引きによって，中国語，日本語ともにおよそ 22%程度の語が翻訳可能であることが示されている．

3.1 実験設定

上記の中英・日英対訳辞書をそれぞれ並列コーパスとみなし，中国語・日本語の辞書項目は形態素解析を適用し，単語に分割した[5]．中英辞書に含まれる中国語の語彙 525,259 語に対してフレーズベース統計的機械翻訳を適用し，日本語訳の候補をスコア付きで上位 10 語まで出力した．なお，本研究においては素性関数の重みとして $\lambda_1 = \lambda_2 = \lambda_3 = 1, \lambda_4 = 3$ を使用した．

フレーズベース統計的機械翻訳における訳語の探索・生成（デコーディング）には Moses[3] と同様の手法を用いた[6]．

3.2 実験結果

中英辞書の中国語 525,259 語（重複を含む）に対して，何らかの日本語訳が得られたものは 385,509 語 (73.4%)，得られなかったものは 139,749 語 (26.6%) であった．これにより，単純な辞書引きに比べて日本語訳が得られる語数を大幅に増やせることが示された．

[2]http://www.wanfangdata.com/
[3]http://qh.library.hb.cn:85/kjxx/yhcb.htm
[4]http://pr.jst.go.jp/others/tape.html
[5]日本語に対しては JUMAN (http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html) を適用した．
[6]語彙翻訳確率および factored model は用いていない．また，同一の翻訳結果の併合，λ_m 等のパラメータの調整は行っていない．

表 2: 日本語訳の例

項目	内容・訳語	スコア	正答
中国語	声 延迟 线存 储器		
英語	acoustic delay line storage		
分野	声 计		
訳 1	音声 遅延 線 記憶 装置	-17.15	
訳 2	音 遅延 線 記憶 装置	-17.51	
訳 3	音声 遅延 記憶 装置	-17.80	
訳 4	音響 遅延 線 記憶 装置	-17.87	○
訳 5	音 遅延 記憶 装置	-18.16	
訳 6	音響 記憶	-18.17	
訳 7	音響 遅延 線 記憶 装置	-18.36	○
訳 8	超 音波 遅延 線 記憶 装置	-18.42	
訳 9	音響 貯蔵	-18.50	
訳 10	音響 遅延 記憶 装置	-18.52	

表 2 に，得られた日本語訳の例を示した．スコアが最も高かった訳は「音声遅延線記憶装置」で，一見正しそうに見えるが，専門用語の翻訳として正しいのは「音響遅延線記憶装置」である[7]．

得られた日本語訳の正確性を評価するために，得られた辞書のうち「计」（コンピュータ）分野から 200 語をランダムに抽出し，人手で正解をチェックした．このうち，何らかの日本語訳が出力されたのは 181 語 (90.5%)，正しい日本語訳が上位 10 語に含まれているものは 135 語 (67.5%)，正しい日本語訳がないものは 46 語 (23.0%) であった．

表 3 は上位 n 個の日本語訳に正しい訳が含まれている割合を表しており，最上位の日本語訳が正しいものは 73 語 (36.5%) であった．また，以下の式で定義される MRR (Mean Reciprocal Rank) は 0.466 であった．

$$\mathrm{MRR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{r_i}, \tag{10}$$

ただし，r_i は i 番目の語に対して日本語訳が現れた最高の順位（ない場合は $1/r_i = 0$）とする．

4 おわりに

本稿では，フレーズベース統計的機械翻訳を用い，日英対訳辞書と中英対訳辞書から英語を介して日中対訳辞書を構成する手法を提案した．評価実験では，専門用語を含む日英・中英対訳辞書を用い，中英辞書に含まれる中国語に対して日本語訳を生成し，その一部を人手で評価した．中英辞書に含まれる中国語のうち，73.4%に何らかの日本語訳を付与することができ，単

[7]本実験では異なる導出による同一の翻訳結果を併合していないため，同一の訳が複数現れている．これらを併合すると，正解訳の順位は向上する可能性がある．

表 3: コンピュータ分野の 200 語の日本語訳（上位 10 語）の評価: Top-n Precision

n	**Top-n precision**
1	36.5% (73/200)
2	47.5% (95/200)
3	54.0% (108/200)
4	58.0% (116/200)
5	60.5% (121/200)
6	61.0% (122/200)
7	63.5% (127/200)
8	65.0% (130/200)
9	67.0% (134/200)
10	67.5% (135/200)

純な辞書引きによる日本語訳（22.2%）に比べ，日本語訳の数を大幅に増やせることが示された．また，コンピュータ分野の 200 語に対して人手で日本語訳を評価した結果，提案手法が最も高いスコアを割り当てた訳語の 36.5%，上位 10 件の訳語の中では 67.5%の訳が正解であった．この水準は辞書としては不十分なものであるが，辞書を人手で作成するための支援，または機械翻訳に用いるための辞書として，応用を検討している．

今後の課題としては，既知の日中対訳辞書を用いた統計的機械翻訳のパラメータ調整，漢字の類似度における個別の漢字を考慮すること，品詞パターンによる単語の並び替えスコアの改良，英語の対訳語彙の活用，および正解の日中対訳辞書を用いた評価，などを考えている．

謝辞　本研究の一部は，文部科学省科学研究費補助金特別推進研究「高度言語理解のための意味・知識処理の基盤技術に関する研究」および科学技術振興調整費・重要課題解決型研究等の推進「日中・中日言語処理技術の開発研究」の助成を受けています．対訳辞書を提供して頂いた北京万方数据股份有限公司 (Wanfang Data Co., Ltd.) 並びに独立行政法人科学技術振興機構に感謝いたします．

参考文献

[1] Francis Bond, Ruhaida Binti Sulong, Takefumi Yamazaki, and Kentaro Ogura. Design and construction of a machine-tractable Japanese-Malay dictionary. In *Proc. of MT Summit VIII*, pages 53–58, 2001.

[2] Chooi-Ling Goh, Masayuki Asahara, and Yuji Matsumoto. Building a Japanese-Chinese dictionary using kanji/hanzi conversion. In *Proc. of the 2nd International Joint Conference on Natural Language Processing*, pages 670–681, 2005.

[3] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics, demonstration session*, pages 177–180, 2007.

[4] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, 2003.

[5] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[6] Kyonghee Paik, Francis Bond, and Shirai Satoshi. Using multiple pivots to align Korean and Japanese lexical resources. In *Proc. of the Workshop on Language Resources in Asia, Natural Language Processing Pacific Rim Symposium 2001*, pages 63–70, 2001.

[7] Charles Schafer and David Yarowsky. Inducing translation lexicons via diverse similarity measures and bridge languages. In *Proc. of the 6th Conference on Natural Language Learning*, volume 20, pages 1–7, 2002.

[8] Satoshi Shirai and Kazuhide Yamamoto. Linking English words in two bilingual dictionaries to generate another language pair dictionary. In *Proc. of 19th International Conference on Computer Processing of Oriental Language*, pages 174–179, 2001.

[9] Kumiko Tanaka and Kyoji Umemura. Construction of a bilingual dictionary intermediated by a third language. In *Proc. of the 15th International Conference on Computational Linguistics*, pages 297–303, 1994.

[10] Masao Utiyama and Hitoshi Isahara. A comparison of pivot methods for phrase-based statistical machine translation. In *Proc. of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 484–491, 2007.

[11] 内山 将夫, 井佐原 均. 統計的機械翻訳におけるピボット翻訳の比較. In **言語処理学会第 *13* 回年次大会発表論文集**, pages 187–190, 2007.

[12] 張 玉潔, 馬 青, 井佐原 均. 英語を介した日中対訳辞書の自動構築. **自然言語処理**, 12(2):63–85, 2005.

[13] 田中 久美子, 梅村 恭司, 岩崎 英哉. 第三言語を介した対訳辞書の作成. **情報処理学会論文誌**, 39(6):1915–1924, 1998.