日中漢字の対応関係の自動獲得と中日語彙翻訳

綱川 隆司† 劉 瀟† 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 / § NaCTeM, UK
{tuna,liuxiao,okazaki,tsujii} at is.s.u-tokyo.ac.jp

1 はじめに

本稿では，中国語から日本語への語彙の翻訳を行うにあたって，漢字の共通性に着目し，それらの間の対応関係・翻訳確率を日中対訳辞書から自動的に獲得することを試みる．また，この翻訳確率を用いて，日英対訳辞書と中英対訳辞書による語彙の統計的翻訳 [17] を行う手法を提案する．

日本語と中国語の間では，同一の漢字の意味が一致または類似しているものが多い．このため，日本語と中国語は語族が異なるにもかかわらず，お互いに文や単語の意味を推測することがある程度可能である．この特徴は，文アラインメント [12]，言語横断情報検索 [3]，および対訳辞書構築 [16, 4] といった日中間の言語処理技術に有用であると考える．

一方，翻訳等の多言語を扱う言語処理においては対訳辞書・対訳コーパス等の資源が不可欠であるが，実際に豊富に存在するのは英語に対する資源であり，日本語・中国語間の直接の対訳資源は英語に比べると規模が小さい．本研究では日英および中英の対訳辞書を用いて中日語彙翻訳を行うことによりこの問題に対処する．翻訳において第三言語を利用する手法，および日中漢字の対応関係を利用する手法はこれまでにも提案されてきた [13, 1, 11, 10, 16, 4] が，本研究ではさらに，上記の日中漢字の対応関係を統計的に導入することでその精度の向上を図る．

2 日中漢字の対応関係

本稿では，日本語と中国語の漢字をコンピュータ上で統一的に扱うために，文字集合として Unicode を用いる．Unicode では，同一表記の文字に共通のコードが割り当てられており，異なる言語の漢字間のマッピングを得ることができる．このマッピングを用いて，中国語の漢字をそのまま日本語の漢字として扱うことで，そのまま中日間の語彙の翻訳できる場合もある．しかし，実際にはそれぞれの言語において漢字の字体が変化したり，漢字の持つ意味が異なったりする場合があり，同一表記のみによる対応付けでは不十分である．

表 1: 新字体・簡体字・繁体字の関係

	新字体	簡体字	繁体字	中国語の意味
1	字	字	字	
2(a)	姫	姬	姬	
2(b)	来	来	來	
2(c)	書	书	書	
2(d)	広	广	廣	
3	走	走	走	歩く
4	湯	汤	湯	スープ

漢字の字体は，中国語のうち現在は台湾・香港等で主に用いられる繁体字を原形とし，中国においては簡体字，日本においては新字体としてそれぞれ独自に簡略化された．従って，簡体字・新字体のうち簡略化されたものについては原形となる繁体字が存在し，それらの変換テーブルがあれば対応付けがほぼ可能である．日本語の新字体，中国語の簡体字・繁体字の関係は，漢字の持つ意味が日中間で異なる場合も含めると，以下のようにまとめられる．

1. 新字体・簡体字・繁体字の字体が一致し，共通する意味がある場合
2. 共通する意味があるが，字体が異なる場合
   (a) 簡体字・繁体字が一致，新字体は異なる場合
   (b) 新字体・簡体字が一致，繁体字は異なる場合
   (c) 新字体・繁体字が一致，簡体字は異なる場合
   (d) 新字体・簡体字・繁体字がすべて異なる場合
3. 新字体・簡体字・繁体字は一致するが，日中間で意味が異なる場合
4. 新字体・簡体字・繁体字が簡略化の変換テーブルで対応付けられるが，日中間で意味が異なる場合

それぞれの例を表 1 に挙げる．

1. および 2. のケースは，変換テーブルが利用可能であれば対応付けが可能になる一方で，4. のケースのために異なる意味の漢字を結びつけるおそれがある．さらに，同じ意味を表すのに日本語と中国語で全く異なる漢字を用いる場合もある[1]．

本研究では，日中対訳辞書から統計的に漢字のアラインメントを得ることで日本語・中国語の漢字を対応付ける．まず日中対訳辞書の項目のうち，日本語が漢字のみで構成されているものを抽出し，各漢字を 1 単語とみなした対訳コーパスとして扱う．これに対して IBM Models [2, 8] を用いて中国語の漢字 c から日本語の漢字 j への翻訳確率 $p(j|c)$ を得ることができる．本研究ではこの翻訳確率を日中間の漢字の対応関係として扱う．ただし，対訳辞書内で共起しない漢字ペア (j, c) の翻訳確率は，$p(j|c) = \varepsilon = 1/N$ （N は漢字の総種類数）とする．中国語の漢字 c に対応する日本語の漢字がない確率も，$p(\mathrm{NULL}|c)$ として同時に得ることができる．

3 中日語彙翻訳

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 [14] を用いて英語を介した中日間の語彙の翻訳を行う．また，自動獲得した日中漢字間の対応関係を語彙翻訳に導入し，その有用性を示す．

フレーズベース統計的機械翻訳のための訓練データとして，日中対訳辞書の他に，日英対訳辞書と中英対訳辞書があると仮定する．各言語の形態素解析を行って単語分割を行い，各対訳辞書を並列コーパスとみなしてフレーズベース統計的機械翻訳 [8] の枠組みを用いて各言語対の対訳フレーズ対と，その翻訳確率を得る．日英および中英の対訳フレーズ対からは，以下の式により英語を介した日中対訳フレーズ対を新たに作成する[2]．

$$p'(\bar{w}_J|\bar{w}_C) = \sum_{\bar{w}_E} p(\bar{w}_J|\bar{w}_E)p(\bar{w}_E|\bar{w}_C), \quad (1)$$

$$p'(\bar{w}_C|\bar{w}_J) = \sum_{\bar{w}_E} p(\bar{w}_C|\bar{w}_E)p(\bar{w}_E|\bar{w}_J). \quad (2)$$

3.1 素性関数

フレーズベース統計的機械翻訳では，対数線型モデルを用いて，以下の式を用いて中国語の用語 $\mathbf{c}$ を日本語の用語 $\hat{\mathbf{j}}$ に翻訳する．

$$\hat{\mathbf{j}} = \operatorname*{argmax}_{\mathbf{j}} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{j}, \mathbf{c}), \quad (3)$$

ただし，$h_m(\mathbf{j}, \mathbf{c})$ は $\mathbf{j}, \mathbf{c}$ 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．

素性関数としては，以下のものを用いた．フレーズ翻訳確率については，日中・中日の両方向を用いた．

- 日中対訳辞書から求めたフレーズ翻訳確率
- 日英・中英対訳辞書から求めたフレーズ翻訳確率
- 日本語言語モデル
- フレーズの並び替えスコア [6]
- 日中漢字の対応関係に基づくスコア

日中漢字の対応関係に基づくスコアは，漢字の翻訳確率 $p(j|c)$ を用いて以下のアルゴリズムを再帰的に実行して求めた．

1. 中国語の最初の文字に対して，漢字の翻訳確率が高い日本語の文字を順に d 個[3]まで求める．
2. 中国語の最初の文字，および対応する日本語の文字（存在する場合）を削除した残りの文字列対の d 個の可能性についてそれぞれ再帰的に 1. を実行し，得られた値と対応する漢字対の翻訳確率の積のうち最大の値を返す．

4 実験

4.1 実験設定

実験に用いたデータは専門用語を含む日英・中英対訳辞書および一般語を含む日中対訳辞書である．中国語は簡体字で記述されている．

中英 万方数据 (Wanfang Data) 英汉-汉英科技大词库[4]: 525,259 項目
日英 JST 機械翻訳辞書[5]: 527,206 項目
日中 EDR 日中英辞書[6]: 596,967 項目

表 2 に対訳辞書に含まれる語彙数を示した．各対訳辞書を並列コーパスとみなし，形態素解析[7]を適用した．

日本語の言語モデルとして，訓練データに含まれる日本語，言語処理学会論文誌の一部等の日本語論文テキスト，および Web 日本語 N グラム [15] から構築した 3-グラム言語モデルを用いた．

翻訳に用いるテストデータとして，中英辞書の一部（医学分野，48,251 項目）に日本語訳を付与し，その中からパラメータ最適化用の開発データとテストデー

[1] 例として，日本語の「歩」と中国語の「走」，日本語の「押」「引」（ドアを押す・引く）と中国語の「推」「拉」など．

[2] $\bar{w}_J, \bar{w}_C, \bar{w}_E$ はそれぞれ日本語・中国語・英語のフレーズを表す．

[3] d はビームサーチのビーム幅を示す．

[4] http://qh.library.hb.cn:85/kjxx/yhcb.htm

[5] http://pr.jst.go.jp/others/tape.html

[6] EDR 電子化辞書 (http://www2.nict.go.jp/r/r312/EDR/J_index.html) の日英対訳辞書の一部に中国語訳を付加したもの．

[7] 日本語に対しては JUMAN (http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html)，中国語に対しては [7] による形態素解析器を用いた．

表 2: 対訳辞書の語彙数

対訳辞書	中国語	英語	日本語
中英	375,990	429,807	-
日英	-	418,044	465,563
日中	233,363	-	110,436

タを 1,000 項目ずつ無作為に抽出した．開発データおよびテストデータの中国語の用語に対して以下の各手法を適用し，日本語の訳語を出力させて評価した．

同一表記 中国語の表記をそのまま日本語の表記と解釈して出力する．

漢字変換テーブル 簡体字から新字体への変換テーブルを適用して出力する．

漢字翻訳確率 中国語の各漢字について，漢字翻訳確率が最も高い漢字をそれぞれ出力する．

中英・日英辞書による辞書引き 中国語の入力語を中英辞書を用いて英語にし，さらに日英辞書を用いて日本語にして出力する[8]．

日中辞書による翻訳 日中対訳辞書のみを訓練データとして用いて翻訳する．

日中・中英・日英辞書による翻訳 すべての辞書を訓練データとして用いて翻訳する．

実験にはフレーズベース統計的機械翻訳のツールキットとして Moses [5] を用いた．

評価指標として，翻訳結果（1-best, 10-best）と正解訳との一致率のほか，mean reciprocal rank (MRR)[9]，BLEU（文字単位，単語単位）[9] を用いた[10]．

4.2 実験結果

表 3 に実験の評価結果を示した．同一表記の結果から，テストセットには中国語と日本語が字体も含めて同一の表記であるものが 6.6%含まれていることがわかる．さらに，簡体字・新字体の変換テーブルを用いることで 11.1%が正しく翻訳できた．この値は漢字の表記のみによるアプローチがある程度有用であることを示しているが，文の翻訳をはじめ多言語の自然言語処理に適用するにはまだ不十分な精度である．日中対訳辞書から求めた日中漢字の翻訳確率を導入することにより，精度の改善がみられた．日英・英中辞書を用いて辞書引きをした場合，同一表記で正解したものも含め一致率が 17.6%に達し，日中辞書のみによる翻訳を大きく上回った．さらに，日英・英中辞書を訓練データに加えることにより，20%を超える一致率を得た．これにより英語を介した手法が日中翻訳に対しても有用であることが示された．

表 3: 評価結果

実験設定	1-best 一致率	文字単位 BLEU
同一表記	0.066	0.1415
漢字変換テーブル	0.111	0.1964
漢字翻訳確率	0.120	0.2347
中英・日英辞書による辞書引き	0.176	0.2774
日中辞書による翻訳	0.123	0.2337
全辞書による翻訳	0.217	0.3438

実験設定	10-best 一致率	MRR	単語単位 BLEU
日中辞書による翻訳	0.144	0.1302	0.0856
全辞書による翻訳	0.272	0.2330	0.2580

表 4 に翻訳結果の例を示した．下線部は正解訳との差分を表している．「膝关节肌」の翻訳においては，中国語の「肌」を日本語の肌（はだ）ではなく筋肉という意味に訳すため，漢字の同一表記および変換テーブルによる手法では翻訳に失敗するのに対し，漢字変換テーブルを用いた手法では正しく「筋」と訳されている．「指浅屈肌」の翻訳でも同様だが，この場合は漢字の順序の入れ替えを伴うために失敗する．「指浅」→「浅指」の翻訳のほか，「突变」→「突然变异」のように文字数が異なるものや，「阑尾」→「虫垂」のように単語単位で翻訳する必要がある場合は，フレーズベース統計的機械翻訳によって翻訳を行うことができた．

5 おわりに

本稿では，日中漢字の対応関係およびそれらの翻訳確率を日中対訳辞書から統計的に求める手法を提案した．また，その翻訳確率を素性の一つとして，日英・中英対訳辞書を用いた英語を介したフレーズベース統計的機械翻訳を行って中国語から日本語への語彙の翻訳実験を行った．評価実験においては，テストセットに含まれる中国語のうち 21.7%について人手で付与した日本語と一致する翻訳結果を得ることができ，漢字の置換，日英・中英辞書を用いた辞書引き，および日中対訳辞書のみを用いた翻訳に比べて高い精度を得ることができた．この水準は，未知語に対して訳語を自動的に生成する実際のアプリケーションとしてはまだ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，または文単位の機械翻訳システムへの導入などが期待

[8] いずれかの辞書に項目が存在しない場合は翻訳せずに同一表記のまま出力．

[9] $\mathrm{MRR} = (1/N)\sum_{i=1}^{N}(1/r_i)$, 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．

[10] 10-best の一致率，MRR，単語単位の BLEU 値についてはフレーズベース統計的機械翻訳を用いた結果のみ示す．

表 4: 翻訳結果の例

入力	下肢	膝关节肌	半致死突变	指浅屈肌	阑尾阻塞
正解訳	下肢	膝関節筋	半致死突然変異	浅指屈筋	虫垂閉塞
漢字変換テーブル	下肢	膝関節 肌	半致死突変	指浅 屈 肌	阑尾阻 塞
漢字翻訳確率	下肢	膝関節筋	半致死 急 変	指薄 屈筋	盲尾支密
中英・日英辞書による辞書引き	下肢	膝関節筋	半致死突 変	浅指屈筋	阑尾阻 塞
日中辞書による翻訳	下肢	膝 关节肌	半 分 致死 急 変 する	指浅 屈筋	盲腸支え
全辞書による翻訳	下肢	膝関節筋	半致死突然変異	浅指屈筋	虫垂閉塞

される．

今後の課題としては，日英・中英・日中の対訳コーパスの利用，名詞句の構文変換パターンの導入，および対訳辞書作成支援等の応用システムに導入した際の有効性の実証が考えられる．

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] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[3] Fredric C. Gey, Noriko Kando, and Carol Peters. Cross-language information retrieval: the way ahead. *Information Processing and Management: an International Journal*, 41(3):415–431, 2005.

[4] 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.

[5] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proc. of the 45th Annual Meeting of ACL, demo. session*, pages 177–180, 2007.

[6] 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.

[7] Tetsuji Nakagawa and Kiyotaka Uchimoto. Hybrid approach to word segmentation and POS tagging. In *Companion Volume to the Proc. of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 217–220, 2007.

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

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. Technical report, IBM Research Division, Thomas J. Watson Research Center, 2001.

[10] 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.

[11] 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.

[12] Chew Lim Tan and Makoto Nagao. Automatic alignment of Japanese-Chinese bilingual texts. *IEICE Transactions on Information and Systems*, E78-D(1):68–76, 1995.

[13] 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.

[14] 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.

[15] 工藤 拓, 賀沢 秀人. Web 日本語 N グラム第 1 版, 2007.

[16] 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.

[17] 綱川 隆司, 岡崎 直観, 辻井 潤一. 日英・中英対訳辞書からの日中対訳辞書の構築. In **言語処理学会第** *14* **回年次大会発表論文集**, pages 464–467, 2008.