日中共通漢字の整理とこれを利用した日中機械翻訳の高度化

京都大学大学院情報学研究科 中澤 敏明

PROFILE 2010年京都大学大学院情報学研究科知能情報学専攻博士課程修了。博士（情報学）。機械翻訳の研究に従事。

nakazawa@nlp.ist.i.kyoto-u.ac.jp TEL 075-753-5346

京都大学大学院情報学研究科 Chu Chenhui

PROFILE 2012年9月京都大学大学院情報学研究科知能情報学専攻修士課程修了。機械翻訳の研究に従事。

京都大学大学院情報学研究科教授 黒橋 禎夫

PROFILE 1994年京都大学大学院工学研究科電気工学第二専攻博士課程修了。博士（工学）。2006年4月より京都大学大学院情報学研究科教授。自然言語処理、知識情報処理の研究に従事。

1 はじめに

近年、中国に出願される特許件数が著しく増加しており、中国語で書かれたこれら特許の閲覧が必要な機会が増えている。しかし英語文献の閲覧に比べて、中国語文献の閲覧は敷居が高い。また逆に中国語でなんらかの文書を書く機会もあるだろう。このような背景から、日中機械翻訳の必要性がより一層高まっているが、高精度な日中機械翻訳システムはいまだに実現には至っていない。日中機械翻訳の高度化の一つの手掛かりとなるのが、漢字情報である。日本語の漢字は中国語に由来するため、多くの漢字を共有している。同じ漢字は同じ意味を持っている場合が多く、多くの日中自然言語処理タスクに有効であり、特に機械翻訳における単語対応の自動推定（アライメント）の精度向上に有効であると考えられる。

おもに中国本土やシンガポールで使われている文字は簡体字（Simplified Chinese）と呼ばれるもので、その名の通り文字が簡略化されており、元の字体とは大きく異なるものが多い。一方で日本語の漢字に近いのは台湾、香港、マカオなどで使われている繁体字（Traditional Chinese）と呼ばれるものであり、簡体字とは大きく異なる。以降、本稿では日本語と簡体字で書かれた中国語にのみ焦点をあてる。

日本語の漢字と簡体字では、現在でも全く同じ形をしているものもあるが、部分的に相違点はあるが同じ漢字だと判断できるものや、見た目では同じかどうか判断できないものが多く存在する。たとえば【表1】に示すように、日本語の「雪」という漢字は中国語でも「雪」と書くが、「愛」は少しだけ異なって「爱」、「発」は大きく異なり、「发」と書かれる。また中国語には存在しない日本語の漢字、いわゆる国字も存在する。

表1：日本語漢字と簡体字の対応例

	同一	違い小	違い大	国字
日本語	雪、安	愛、詞	発、広	込、畑
簡体字	雪、安	**爱、词**	**发**、广	なし

本稿では、既存のさまざまなリソースを活用して日本語漢字と簡体字のマッピングテーブルを作成し、これを用いて日中アライメントを高度化する方法を提案する。

2 漢字マッピングテーブル

漢字マッピングテーブルは、日本語で使われる漢字を基として、簡体字を対応付けることで構築する。しかしどの文字を日本語の文字とするのかも難しい問題である。ここではJIS漢字コードの第1、第2水準漢字、

合わせて 6355 文字を日本語の漢字として用いる [2]。

2.1　漢字マッピングテーブルの構築

構築には以下の３つのリソースを利用した。１つ目は Unihan database[7] である。これは Unicode Consortium が構築している、CJK 統合漢字に関するデータベースであり、全ての Unicode 漢字についての情報が含まれている。そこには“mappings”、“readings”、“dictionary indices”、“radical stroke counts”、“variants”など様々な情報が含まれているが、このうち日本語と中国語の漢字の関係を示しているのは“mappings”と“variants"であるので、この二つを利用した。

２つ目は中国語の繁体字を簡体字に変換するオープンソースツール Chinese encoding converter[8] である。このツールには 6740 の対応する繁体字と簡体字のペアが収録されている。前述の通り、日本語の漢字は繁体字に近いので、このツールを利用することにより、繁体字を通して間接的に日本語漢字と簡体字との対応をいくつか発見することができる。

３つ目は Kanconvit[9] である。このツールは日本語の漢字と簡体字間の変換を行うことができるフリーのツールであり、3506 の１対１対応テーブルを用いている。しかし実際には、漢字の対応は１対１でない場合もある。たとえば日本語の「発」と「髪」は簡体字ではどちらも「发」である。

これら３つのリソースには互いに共通する情報もあるが、各リソースにしか存在しない情報もある。そこで我々は、これら全てを統合・整理し、網羅的な漢字マッピングテーブルを構築する。マッピングテーブルは、日本語漢字と簡体字で同一のもの、異なるもの、日本語にしか存在しない（対応する簡体字がない）ものの３つのカテゴリーからなる。なお日本語にしか存在しないものには、国字である場合と３つのリソースに情報がなく発見できなかった場合の２種類がある。

2.2　マッピングテーブルの性質と網羅性

構築したマッピングテーブルの各カテゴリーに含まれる文字数の変化を【表2】に示す。Unihan database だけでもかなりの文字の対応が発見できるが、その他二つのリソースを追加することにより、新たにいくつかの対応が発見できていることがわかる。

表２：マッピングテーブル内の各カテゴリーの文字数の変化

	同一	相違	対応なし
Unihan	3318	2364	673
+ converter	3318	2401	636
+ Kanconvit	3318	2413	624

また日中科学技術論文コーパスを用いて、構築したマッピングテーブルの網羅性を調査した。このコーパスは、科学技術振興調整費による重点課題解決型研究「日中・中日言語処理技術の開発研究」において構築されたものであり、約 68 万対訳文からなる。このコーパス内の日本語文に含まれる漢字は約 1400 万文字、中国語文の漢字は約 2400 万文字である。このうち、同一の文字（Unicode が共通の文字）は日本語で 52.41%、中国語で 30.48% であった。さらに、構築したマッピングテーブルを用いて共通漢字を考慮すると、日本語で 76.66%、中国語で 44.58% をカバーすることができた。日本語文に含まれる文字の４分の３以上は、中国語に共通する漢字があるということがわかり、いかに漢字情報が強力な情報源であるかがわかる。

3　共通漢字情報のアライメントでの利用

共通する漢字は同じ意味を持っていることが多い。このため、同じ内容が書かれている日中対訳文において、共通漢字が出現する可能性や割合は高いと考えられる。これは前章の実験からも明らかである。そこで共通漢字情報を機械翻訳、特に対訳文内の単語対応の自動推定に利用することを考える [1]。

3.1　アライメントモデル

アライメントモデルとして、単語依存構造木を利用したベイジアン部分木アライメントモデルを利用する

[5]。詳細は省略するが、このモデルには部分木ペアを生成する確率、言い換えれば、２つの部分木が対応関係にある確率が使われており、共通漢字情報を利用してこの確率を修正することで、正しい対応を発見しやすくする。具体的には、元の部分木ペア生成確率を$\theta_T(\langle e, f \rangle)$とすると、重み w を用いて$w \cdot \theta_T(\langle e, f \rangle)$と修正する。

重み w は各部分木に含まれる漢字のうち、共通漢字の割合を基に計算される。たとえば日本語の「実際」という単語と中国語の「事実」という単語では、「実 ⇔ 実」という共通漢字があるため、割合は 2/4=0.5 となる。重み w は共通漢字の割合に 5000 を掛けた値を用いる。ただし、ここで注意が必要なのは、単純に 5000 という値を掛けているため、全体として確率値ではなくなってしまっていることである。全体を確率的なモデルにしたまま共通漢字情報を用いるのは今後の課題である。

３.２　アライメント実験

共通漢字情報を利用することの有効性を示すために、前章と同じ日中科学技術論文コーパスを用いてアライメント実験を行った。アライメントの正解データとして、人手で正解が付与された 510 文を利用した。このデータには Sure（必須の対応）と Possible（あっても間違いではない対応）の２種類の正解が付与されている。精度は Precision、Recall、Alignment Error Rate (AER) を用いて評価した。AER はアライメントの総合的な精度を示す指標であり、その値が低い程精度が良い。

$$\text{Precision} = \frac{|A \cap P|}{|A|} \quad \text{Recall} = \frac{|A \cap S|}{|S|} \quad \text{AER} = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

A がシステムの出力、P と S が正解である。比較として、GIZA++[4] と BerkeleyAligner[3] によるアライメントも行った。また baseline は共通漢字情報を利用しないモデルである。

実験結果を【表３】に示す。baseline は GIZA++ や BerkeleyAligner よりも低い AER を示しているが、共通漢字情報を利用することによりさらにアライメント誤りを低減することができた。【図１】にアライメント精度が向上した例を示す。baseline では「規準」と「規定」の対応が発見できていないが、「規 ⇔ 規」の共通漢字情報を利用することで、提案手法では対応を発見することができた。

表３：アライメント実験結果

	Pre.	Rec.	AER
GIZA++	83.77	75.38	20.39
Berkeley	88.43	69.77	21.60
Baseline	85.37	75.24	19.66
Proposed	85.55	76.54	18.90

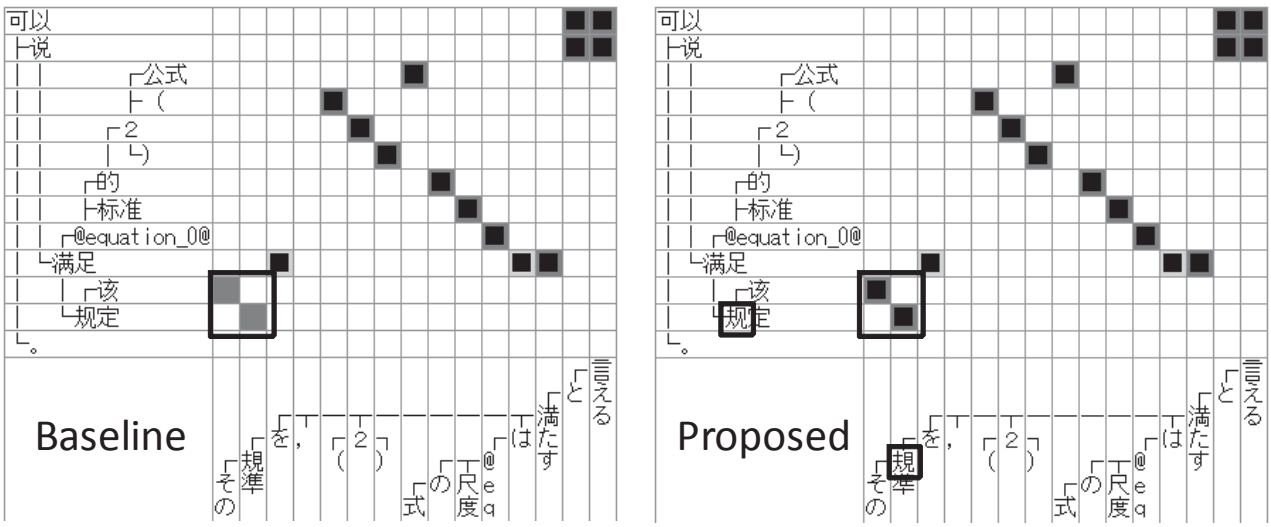

図１：アライメントの改善例

3.3　翻訳実験

京都大学で開発している用例ベース機械翻訳システム[6]を用いて翻訳実験を行った。トレーニングに用いたコーパスはアライメント実験のものと同じである。テスト文は同じ論文ドメインの1770文を用いた。評価は翻訳の自動評価指標であるBLEUを用いた。【表4】に結果を示す。わずかではあるが、翻訳精度が向上していることがわかる。共通漢字情報によりアライメント精度が向上したことにより、質の高い翻訳知識を構築することができ、結果的に翻訳精度も向上したものと考えられる。

表4：翻訳実験結果（BLEU）

	日→中	中→日
Baseline	22.84	19.10
Proposed	23.14	19.22

4　まとめと今後の課題

本稿では、日中共通漢字マッピングテーブルの構築および、これを利用した日中機械翻訳の高度化について述べた。日中共通漢字情報は、機械翻訳そのものの高度化はもちろんのこと、コンパラブルコーパスからの訳語、対訳文抽出などにも利用可能である汎用的な知識である。今後はこれらのタスクでの有効利用も検討する。機械翻訳での利用に関しての課題は、1つは前にも述べたように、より精錬された方法でアライメントモデルに組み込むことである。また、ある単語のペアが共通漢字を含んでいるからといって、どんな時でも同じ意味を持っているかといえばそうではない。このような場合は共通漢字情報が悪影響を与えることがある。この問題に対処するために、文脈等を考慮して共通漢字の妥当性を判断することも今後検討する。

参考文献

[1] Chenhui Chu, Toshiaki Nakazawa and Sadao Kurohashi: Japanese-Chinese Phrase Alignment Using Common Chinese Characters Information, Proceedings of MT Summit XIII, pp. 475-482, 2011
[2] Chenhui Chu, Toshiaki Nakazawa and Sadao Kurohashi: Chinese Characters Mapping Table of Japanese, Traditional Chinese and Simplified Chinese, In Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC'12), 2012
[3] John DeNero and Dan Klein: Tailoring Word Alignments to Syntactic Machine Translation, Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, pp. 17-24, 2007
[4] Philipp Koehn, Franz J. Och and Daniel Marcu: Statistical Phrase-based Translation, HLT-NAACL 2003: Main Procedings, pp. 127-133, 2003
[5] Toshiaki Nakazawa and Sadao Kurohashi: Bayesian Subtree Alignment Model Based on Dependency Trees. In Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP2011), pp. 794–802, 2011.
[6] Toshiaki Nakazawa and Sadao Kurohashi: EBMT System of KYOTO Team in PatentMT Task at NTCIR-9, Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-9), 2011
[7] http://unicode.org/charts/unihan.html
[8] http://www.mandarintools.com/zhcode.html
[9] http://kanconvit.ta2o.net/