日中特許対訳コーパスを用いた対訳辞書の自動構築

安田　圭志　　隅田　英一郎
情報通信研究機構
{keiji.yasuda,eiichiro.sumita}@nict.go.jp

1　はじめに

近年、中国国内における特許出願数は大幅な伸びを見せており、2011 年には 50 万件を突破した。この数は日本国内における出願数の 1.5 倍である。このようなことから、現在、中日特許分野における、機械翻訳、多言語情報検索等に代表される、自然言語処理技術の開発が、益々重要になってきている。

これらの自然言語処理技術においては、対訳辞書は非常に重要な役割をはたしているが、本研究では、対訳辞書を自動構築する方法を提案する。従来から、種々の言語対において、対訳辞書自動生成の研究が行われているが、その多くは、既存の対訳辞書から得られる情報を用いて、辞書のエントリーを拡張するという方法 [2, 8, 5] がとられている。

本研究では、ベースとなる対訳辞書を必要とせず、日本語漢字から中国語簡体字への文字マッピング知識と、既存の統計的機械翻訳、用語抽出等の各種自然言語処理ツールとを用いて、日中特許対訳コーパスから対訳辞書を段階的に自動構築する。

2　提案手法

提案手法では、日本語漢字と中国語簡体字との置換や、種々の自然言語処理ツールを用いる。図 1 に示すように、提案手法は 3 つのステップからなる。まず、前処理として、茶筌 [1] と ICTCLAS [9] により、日中対訳コーパスの形態素解析を行う。日本語形態素解析結果は、用語抽出システム [6] に渡される。提案手法では、ここで抽出された用語に対する対訳を、対訳コーパスの中国語側を用いることにより生成する。

提案手法の処理の流れを図 2 と図 3 に示す。各ステップの詳細については、次節で述べる。

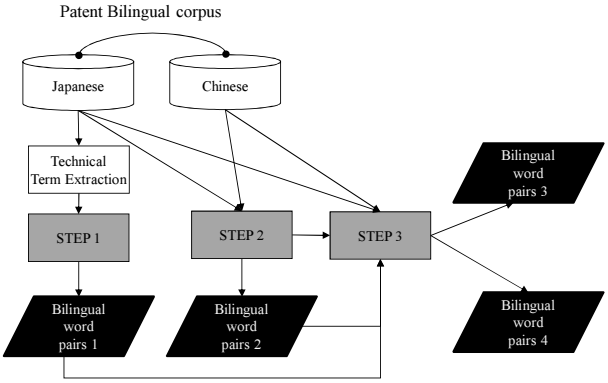

図 1: 提案手法の概要

2.1　STEP1: 漢字置換

STEP1 では、日本語漢字と中国語簡体字との置換処理を行う。ここでは、前処理で抽出された日本語用語の内、漢字のみからなる用語に対して処理を行う。単純に漢字を簡体字に置換した場合、中国語の単語として正しい用語と、誤った用語が生成される。提案手法では、中国語の単語としての正しさを担保するため、生成された簡体字文字列が、対訳コーパスの中国語側に出現するかどうかのチェックを行い、中国語側に出現した場合のみ、日本語用語の対訳として採用する。STEP1 で抽出された訳語対は、図 2 の、Bilingual word pairs 1 に対応する。

2.2　STEP2: フレーズテーブルの利用

STEP2 では、2 種類の異なる統計的機械翻訳学習ツールを用いる。まず、MOSES [3] と pialign [7] とを用い、日中対訳コーパスから 2 種類のフレーズテーブル生成する。次に、これらの 2 つのフレーズテーブルから、共通するフレーズを抽出することにより、精度の高い訳語対を生成する。STEP2 で抽出された訳語対は、図 2 中の Bilingual word pairs 2 に対応する。

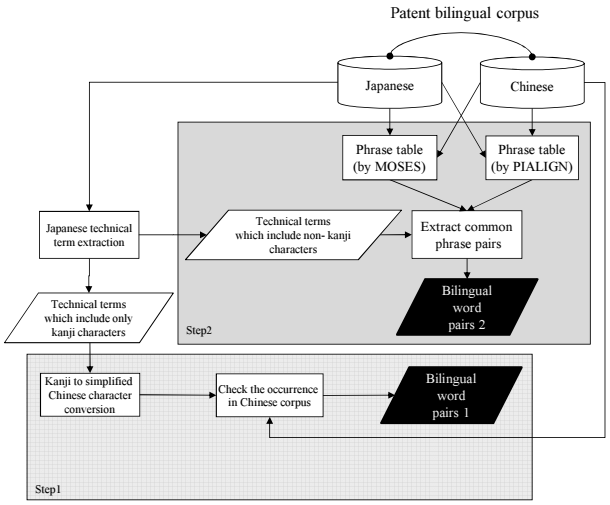

図 2: 提案手法の枠組み (Step 1 〜 2)

2.3 STEP3: 統計的機械翻訳の利用

STEP3 においては、統計的機械翻訳の枠組みを用いる。提案手法では、学習データの違いにより 2 つの単語翻訳システム（SMT 1 と SMT 2）を構築する。SMT 1 は、日中対訳コーパスを全て用いて学習された翻訳システムである。SMT 2 は、STEP1 と STEP2 で得られた訳語対のみを用いて学習された翻訳システムである。SMT 2 では、一般の統計的機械翻訳における文の単位を複合語に、統計的機械翻訳における単語の単位を複合語を構成する部分単語に置き換えて学習処理を行なう。また、両システムとも、分割された日本語の複合語を入力とし、モノトニックデコーディングにより翻訳し、中国語単語を出力する。2 システムによる出力は、図 4 に示された方法により選択される。

図 4 に示すように、SMT 2 の出力の内、SMT 1 の出力と一致した場合については、Bilingual word pairs 3 とし、一致しない場合については、Bilingual word pairs 4 とする。このように、SMT 1 は、SMT 2 の出力の信頼度を見るために、補助的に用いられており、SMT 1 の出力が、直接対訳として用いられることは無い。

3 実験

3.1 実験条件

実験では、Lu ら [4] により構築された、993 K 文対からなる日中対訳コーパスを用いた。前処理における用語抽出では、日本語用語 603 K 単語が抽出され

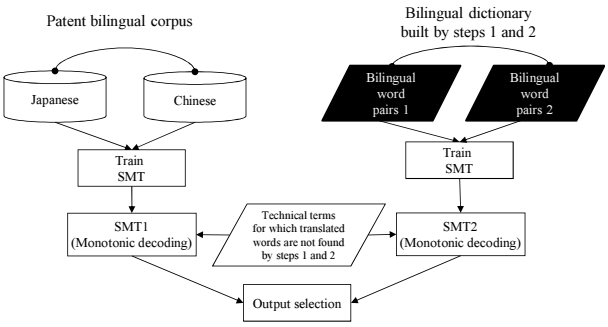

図 3: 提案手法の枠組み (Step3)

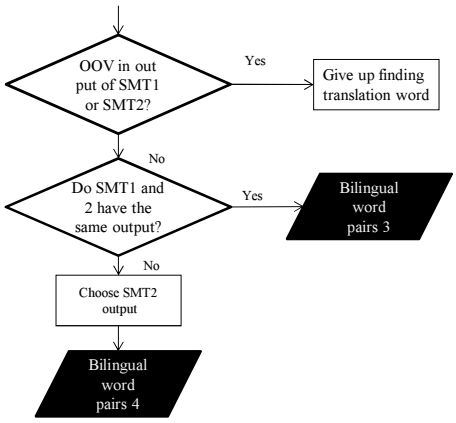

図 4: STEP3 における選択処理の流れ

た。この内の約 40%が、STEP1 における処理対象となる漢字のみからなる用語であった。残りの約 60%が、STEP2 により処理される。STEP1 および STEP2 において、対訳が得られなかった日本語用語については、STEP3 で再度処理される。

3.2 実験結果

表 2: 提案手法の評価結果 (テストセット precision)

Type of extracted word pairs	Test set precision
Bilingual word pairs 1	92.1%
Bilingual word pairs 2	79.5%
Bilingual word pairs 3	85.6%
Bilingual word pairs 4	59.3%

表 1 に提案手法の各 STEP により得られる訳語対の数を示す。STEP3 までで得られる Bilingual word pairs 1〜4 を全て含めると、92. 8%の日本語用語に対して何らか訳語が生成されている。

表 2 は、各ステップで得られる訳語対の precision である。ここでの評価では、各訳語対のグループ毎に 200 単語対づつ無作為抽出したテストセットを用いて、

表 1: 提案手法により抽出された訳語対の数

		Type of Japanese techinical term		Total	Percentage	Accumulative percentage (Yield ratio)
		Kanji only	Term including non-kanji character			
Type of extracted word pairs	Bilingual word pairs 1	58,951	N/A	58,951	9.8%	9.8%
	Bilingual word pairs 2	N/A	136,061	136,061	22.6%	32.3%
	Bilingual word pairs 3	43,318	80,471	123,789	20.5%	52.9%
	Bilingual word pairs 4	130,951	109,813	240,764	39.9%	92.8%
No translation extracted		8,675	34,641	43,316	7.2%	100.0%
Total		241,895	360,986	602,881		

日中バイリンガルの評価者による、主観評価を実施している。評価基準は、生成された中国語単語が、日本語用語の対訳として完全に正しいかどうかである。なお、前処理の日本語用語抽出部での誤りが生じている訳語対については、評価から除外している。

表 1 と表 2 に示したように、Bilingual word pairs 1 における precision は 92.1%と高いが、この段階で訳語が得られる用語は、全体の 9.8%のみである。STEP2〜3 で得られる Bilingual word pairs 2〜3 では、前述の Bilingual word pairs 1 よりも precision で劣るものの、80%前後の値が得られている。Bilingual word pairs 4 の precision は 59.3%と最も低く、Bilingual word pairs 3 よりも 10.7 ポイント低くなっている。また、Bilingual word pairs 4 のテストセットを用いて、SMT 1 の出力を評価した所、precision は、57.8%であり、提案手法よりも 1.5 ポイント低くなった。このことより、図 4 で示した SMT 1 の補助的利用法が適正に作用していることが分かる。

次に、表 1 と表 2 で示した結果を用い、次式により、Bilingual word pairs 1〜4 を結合した際の、全体の precision と recall を推定した。

$$P_1^n = \frac{\sum_{i=1}^{n} t_i \times w_i}{\sum_{i=1}^{n} w_i} \quad (1)$$

$$R_1^n = \sum_{i=1}^{n} t_i \times w_i \quad (2)$$

ここで P_1^n と R_1^n は、Bilingual word pairs 1〜n を全て結合した場合の、precision と recall をそれぞれ表す。また、t_i は、表 2 に示した Bilingual word pairs i のテストセットにおける precision である。w_i は、表 1 に示した Bilingual word pairs i の percentage の値を表しており、ここでは、重みとして利用している。

式 (1〜2) により得られた、precision と recall の推定値を表 3 に示す。Bilingual word pairs 1〜4 すべてを結合すると、recall と precision の推定値がそれぞれ、68.2%、73.5%となった。

表 3: 提案手法の評価結果 (precision と recall の推定値)

Type of extracted word pairs	Estimated precision	Estimated recall
Bilingual word pairs 1	92.1%	9.0%
Bilingual word pairs 1 to 2	83.3%	26.9%
Bilingual word pairs 1 to 3	84.2%	44.5%
Bilingual word pairs 1 to 4	73.5%	68.2%

3.3 STEP3 での選択手法に関する議論

ここでは、STEP3 における選択手法に関する議論を行うため、Bilingual word pairs 3 と 4 の評価結果の詳細分析を行なう。前述の Bilingual word pairs 3 および 4 におけるテストセットを、日本語の漢字のみからなる用語と、それ以外の用語に分類し、それぞれについて、SMT 1 と SMT 2 の出力に対する評価結果を集計した。

表 4 に示すように、日本語側の分類により precision の値に大きく差がでている。漢字のみの入力の場合では、Bilingual word pairs 3 の precision は 90%近くの値を示している。一方、漢字以外を含む入力の場合では、7.8 ポイント程度の劣化が生じている。また、Bilingual word pairs 4 では、SMT 1 と SMT 2 の優劣が逆転している。

これらの結果は、サイズの小さいテストセットでの結果であるが、今後、入力の特性を考慮することにより、より優れた選択を行える可能性があると言える。

4 まとめと今後の検討課題

日中対訳コーパスから、対訳辞書を自動抽出する方法を提案した。

提案手法では、各種自然言語ツールと、日本語漢字から中国語繁体字への置換知識を用いることにより、日中対訳コーパスから日中対訳辞書を段階的に自動生成する。

表 4: Bilingual word pairs 3〜4 の詳細な評価結果

Type of Japanese techinical term	Output type	SMT1	SMT2
Kanji only	Bilingual word pairs 3	89.8%	
	Bilingual word pairs 4	53.5%	46.5%
Term including non-kanji character	Bilingual word pairs 3	81.0%	
	Bilingual word pairs 4	60.2%	72.4%

実験では、ランダムに抽出した 200 語からなるテストセットを用いた評価を行った。実験の結果、最大 92.8%の日本語用語に対して何らかの中国語訳が生成された。precision は、生成過程の段階により、59.3%〜92.1%の値となった。

現在、提案手法により、抽出された Bilingual word pairs 1 と 2 について、人手による修正作業を行っており、整備完了後に公開の予定である。今後、本研究で得られた日中対訳辞書が、日中特許に関連した自然言語処理技術の研究開発の一助となれば幸いである。

参考文献

[1] Masayuki Asahara and Yuji Matsumoto. Extended models and tools for high-performance part-of-speech tagger. In *Proceedings of COLING*, pp. 21–27, 2000.

[2] Chooi-Ling Goh, Masayuki Asahara, and Yuji Matsumoto. Building a japanese-chinese dictionary using kanji/hanzi conversion. In *Proceedings of International Joint Conference on Natural Language Processing (IJCNLP)*, pp. 670–681, 2005.

[3] Hieu Hoang and Philipp Koehn. Design of the moses decoder for statistical machine translation. In *Proceedings of ACL Workshop on Software engineering, testing, and quality assurance for NLP*, pp. 58–65, 2008.

[4] Bin Lu, Benjamin K. Tsou, Tao Jiang, Oi Yee Kwong, and Jingbo Zhu. Multilingual patents: An english-chinese example and its application to smt. In *Proceedings of the 1st CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2010)*, 2010.

[5] Yohei Morishita, Liang Bing, Takehito Utsuro, and Mikio Yamamoto. Estimating translation of technical terms based on phrase translation table and a bilingual lexicon (in japanese). *IEICE TRANSACTIONS on Information and Systems*, Vol. J-93D, No. 11, pp. 2525–2537, 2010.

[6] Hiroshi Nakagawa and Tatsunori Mori. Automaic term recognition based on statistics of compound nouns and their components. *Terminology*, Vol. 9, No. 2, pp. 201–209, 2003.

[7] Graham Neubig, Taro Watanabe, Eiichiro Sumita, Shinsuke Mori, and Tatsuya Kawahara. An unsupervised model for joint phrase alignment and extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 632–641, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[8] Takashi Tsunakawa, Naoaki Okazaki, and Jun'ichi Tsujii. Building a bilingual lexicon using phrase-based statistical machine translation via a pivot language. In *Proceedings of the 22nd International Conference on Computational Linguistics Companion volume Posters and Demonstrations*, pp. 127–130, 2008.

[9] Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. Hhmm-based chinese lexical analyzer ictclas. In *Proceedings of the second SIGHAN workshop on Chinese language processing*, Vol. 17, pp. 184–187, 2003.