RBMT、SMT、Hybrid MT の特徴比較と今後の展望

真下 修三　　高 京徹　　趙 東柱　　川上 健
株式会社高電社　　kohk@kodensha.jp

1. はじめに

機械翻訳がこの世に登場してから、そのほとんどがルールベース（RBMT）により開発されていた。昨今は、インターネットの普及と共に大量の対訳コーパスを入手することが可能となり、統計的手法を用いた翻訳（SMT）が普及している。学会等で RIBES や BLEU による評価結果を基に SMT の優位性が強調されることも多い。本稿では RBMT と SMT の特徴や翻訳結果の比較、通信を必要としない RBMT の開発事例のほか、複数の MT の長所を融合させた Hybrid MT の効果や将来について言及する。

2. SMT、RBMT、Hybrid MT の特徴

2.1 SMT の特徴

１）　翻訳対象となる文章に合致する対訳コーパスが存在する場合は、翻訳結果が良くなる傾向がある。
２）　SMT に必要な言語モデル・翻訳モデルのデータは容量が大きくなる[1]ため、現在の携帯端末に組み込んで使うことが現実的でない。
３）　おのずとネットワークに接続することが求められるため、海外では高額なローミングや有償の Wi-Fi に費用を支払うことを想定する必要がある。
４）　実務面での直近の課題としては、良質の対訳コーパスを効率的に収集することであり、さらに法律面ではそれらの対訳コーパスが著作権に抵触しないことが保証される必要がある。

特に、海外から来日する旅行者が日本で Wi-Fi に接続することの不便さを指摘するなど、外国人観光客が国内の通信環境を利用して SMT で意思疎通を図るためには、解決されるべき課題が残されている。国内の都市部においても、通信環境は場所、建物の構造、基地局からの距離に影響を受けており、この点に関しては海外からの旅行者に限らず、国内の居住者も日常的に直面する問題である。

2.2 RBMT のメリット

開発手法によっては容量を小さくすることができるため、ハードウェアへのインストールが容易で、通信を必要としない仕様にすることも可能。

2.3 Hybrid MT のメリット

RBMT と SMT の双方のメリットを活用する手法を、広義で Hybrid MT と呼んでいる。弊社では、特定分野については SMT、その他については RBMT で翻訳する方式を採用している。具体的なメリットとして、ハードウェアリソースやネットワーク環境の許容する範囲内で、対訳コーパスを柔軟に追加できることが挙げられる。たとえば、データセンターはもとより企業内のサーバーのような環境下では、業務に即した膨大な量の対訳コーパスや翻訳メモリを登録し、リソースに制限がある環境下では登録する対訳コーパスの質と量を加減することにより、翻訳結果の最適化を図ることができる。[2]

2.4 SMT と RBMT の翻訳比較

原文		京都的四季**散发着**不同的魅力。		
日本語の語順への並べ替え		京都 的 四季 不同 的 魅力 散発 着。		
理想的な訳文例		京都の四季は異なる魅力を放っている。		
動詞(散发着)の訳出	訳文	MT 種別（自社調べ）	長所	短所
醸し出している	京都の四季さまざまな魅力を醸し出しています。	SMT	動詞の訳出がこなれている	主語の助詞が欠如
			「着」を現在形に訳出している	
	京都異なる季節の魅力を醸し出しています。	SMT	動詞の訳出がこなれている	「四季」の訳出が欠如
			「着」を現在形に訳出している	主語の助詞が欠如
放つ	京都市の四季は違った魅力を放つ。	SMT	主語の助詞を訳出している	「着」を現在形に訳出していない
配布している	京都の四季は違った魅力を配布しています。	RBMT	主語の助詞を訳出している	動詞の訳出が不自然
			「着」を現在形に訳出している	
配っている	京都の四季異なる魅力を配っています。	RBMT	「着」を現在形に訳出している	主語の助詞が欠如
				動詞の訳出が不自然
	京都の四季に異なる魅力を配っている。	RBMT	主語の助詞を訳出している	主語の助詞が誤り
			「着」を現在形に訳出している	動詞の訳出が不自然
拡大されている	京都の四季が拡大されているとは違った魅力的なのだ。	SMT	主語の助詞を訳出している	理想の翻訳結果と大きく異なっている
				訳文に中国語が混在
送り出している	京都の４季節は別の魅力を送り出している。	不明	主語の助詞を訳出している	「四季」を訳せていない
			「着」を現在形に訳出している	動詞の訳出が不自然
撒く	京都の四季は異なっているアトラクションを撒く。	不明	主語の助詞を訳出している	「着」を現在形に訳出していない
				特殊な意訳に近い

原文出典　http://asahichinese.com/category/travel/destination/kyoto/

表中では、9 種類の異なる翻訳エンジンを用いて翻訳結果の検証を行った。スペースの都合上、本稿では比較的に短い原文を紹介したが、上記の翻訳比較では訳文の適確性と流暢性において一定の相違を見出すことができる。特に「魅力」という名詞と「醸し出す」や「放つ」からは組み合わせの妙を感じることができ、コロケーションを参照した SMT ならではの訳出と考えられる。

ただし、異なる原文セットを用いたケースにおいて、特定の翻訳エンジンが常にすぐれた訳出をすることを保証するものではない。原文の長短、分野や構文により、RBMT と SMT とがほぼ同一のレベル、ほぼ同一の訳出をすることがあることは興味深く、筆に値する。

3. RBMT アプリの開発事例

以下に、無通信 RBMT の例として Android および iOS に対応した弊社アプリを紹介する。

3.1 Android 版の画面例

図 1: 中国語を入力して日本語に翻訳した例

図 2: 文字が入力されるたびに翻訳されるインクリメンタル翻訳画面と、ターゲット言語リスト

3.2 仕様と翻訳速度（日中韓ともに、原文には日本の観光情報約 400 文字を採用）

OS	iOS 8				Android 4.4.2					
CPU	Apple A8				MSM8974AB 2.3GHz (quad-core)					
翻訳方向	日→中	中→日	日→韓	韓→日	日→中	中→日	日→韓	韓→日	中→韓	韓→中
ディスク容量（MB）	80	29	96	151	96	35	88	163	137	268
メモリ容量（MB）	12	29	26	27	3	8	3	3	9	9
速度（秒）	0.37	0.28	0.30	0.39	0.72	0.85	0.38	0.37	1.14	0.82

※　中韓および韓中翻訳は、日本語を媒体とするピボット翻訳。

上述のとおり、約 400 文字を翻訳する場合、中韓を除くすべての方向において 1 秒以下で完了することが観測された。1 秒という数値は、オフィス内、国内外の街中を問わずストレスを体感するレベルではなく、通信を利用しない RBMT が翻訳精度と速度の両面において、実用に耐えうることの布石となる。ハードウェアである CPU やメモリの小型化・高速化・低価格化の恩恵を受ける一方で、いかにメモリに負担をかけない小容量の RBMT を開発するかは、通信に頼らない MT を開発する上での醍醐味とも言える。さらに、表中の携帯端末に標準で搭載されている音声認識・合成ソフトウェアと組み合わせることにより、原文入力の手間も簡素化される。結果として、IT 機器の操作に不慣れであっても、通信が寸断された非常事態下にあっても、対面での多言語コミュニケーションに重宝するであろう。

4. 今後の展望

これまでに述べたとおり、RBMT には一定の需要があると推測される。しかし、ディスクやメモリの小型化・大容量化・低価格化は顕著であり、近未来には 1TB 単位の対訳コーパスを携帯端末に保存することさえ想像に難くない。果たしてその時代が到来したときに、RBMT に活路はあるのだろうか？その問いには、次の一文をもって解に代える。

MT 開発の現場において、小は大を兼ねる。

ハードウェアが大型で高額だった時代に開発された小容量の RBMT は、時代の変化とともに潤沢となるハードウェアリソースと結びつき、RBMT と SMT の双方のメリットを最大限に活用できる Hybrid MT へと進化を続ける。

参考文献

[1] NTT コミュニケーション科学基礎研究所　英中韓から日本語への特許文向け統計翻訳システム
須藤 克仁、鈴木 潤、秋葉 泰弘、塚田 元、永田 昌明
[2] 株式会社富士通研究所　富士 秀、鄭 育昌ほか　中日・英日翻訳への定型利用翻訳技術の適用