特許翻訳における機械翻訳システムの評価 ― NTCIR-9特許機械翻訳タスクでの分析 ―

後藤 功雄[†]　Bin Lu[†††,††]　Ka Po Chow[††]
隅田 英一郎[†]　Benjamin K. Tsou[††,†††]

[†] 情報通信研究機構
[††] Hong Kong Institute of Education
[†††] City University of Hong Kong

1 はじめに

特許翻訳は，外国語で書かれた特許の理解や外国への特許出願のために，実用上において大きなニーズがある．そのために特許の機械翻訳の研究は重要であり意義がある．この研究の促進のために評価型ワークショップであるNTCIR-9特許機械翻訳タスクを主催した[4]．NTCIR-9特許機械翻訳タスクでは，特許の機械翻訳の研究に利用できる共通のデータの構築，最新の機械翻訳システムの評価，自動評価手法の信頼性の評価を実施した．このタスクは過去2回の特許翻訳タスク[2, 3]を基にしている．今回は新たに中英翻訳とacceptability評価も行った．本稿では，NTCIR-9特許機械翻訳タスクの概要，評価手法，翻訳システムと自動評価の評価結果について述べる．

2 特許機械翻訳タスクの概要

特許機械翻訳タスクでは，共通のデータを用いて複数の参加グループのシステムおよび主催者によるベースラインシステムを評価することで，特許翻訳におけるシステムの比較評価を可能にするとともに，特許機械翻訳の実用に向けての評価を実施した．

タスク実施の流れは次のとおりである．はじめに主催者が特許文からなる訓練データおよびテストデータを参加グループに提供する．参加グループはそれぞれの手法でテストデータを機械翻訳して提出する．主催者は提出された翻訳結果を評価して，評価結果を参加グループへ返送する．最後にワークショップで参加グループが研究成果を発表する．

実施した翻訳の言語対および翻訳方向は，日本語から英語（日英），英語から日本語（英日），中国語から英語（中英）である．訓練データとして，日英・英日翻訳は約320万文対の日英対訳コーパス，中英翻訳は100万文対の中英対訳コーパス，翻訳先言語の単言語コーパスとして，日英・中英翻訳には英語の特許文3億文以上，英日翻訳には日本語の特許文4億文以上を提供した．開発データには2,000文対，テストデータには2,000文を用いた．対訳コーパス，開発データ，テストデータは，請求項の文を含まず，主に発明の詳細な説明部分の文からなる．訓練データは2005年以前，開発データおよびテストデータは2006〜2007年の特許から構築した．

テストデータは自動的に抽出した対訳文対から，人手で正しい対訳の文対を選択して構築した．テストデータは対訳文対として自動抽出された文から構築しているため，実際の特許文の傾向が完全には反映されていない．この点は次のNTCIR-10[*1]での改善を検討中である．

参加グループ数は全体で21，日英翻訳は12，英日翻訳は9，中英翻訳は18であった．タスクに参加した翻訳エンジンの種類は，大きく分類すると統計翻訳（SMT），ルールベース翻訳（RBMT），用例翻訳（EBMT），複数手法の組み合わせ（HYBRID）の4種類である．さらに，主催者がベースラインシステムの結果として日英・英日・中英翻訳に対して2種類のSMT（フレーズベースSMT，階層フレーズベースSMT）[*2]とGoogle翻訳，日英・英日翻訳に対して3つの商用RBMT，中英翻訳に対して2つの商用RBMTによる翻訳結果を追加した．

3 評価手法

人手による文単位の訳質評価を実施した．各システムあたり3人の評価者がそれぞれ100文，合計300文を評価した．評価基準として，adequacyとacceptabilityの2つを用いた．以下に本評価で用いたこれらの評価基準について説明する．

3.1 Adequacy

翻訳の適切さ（adequacy）の評価を5段階（1-5）で実施した．この評価の目的は，システム間の比較である．本評価では，節レベルの訳質までを考慮して評価した．

3.2 Acceptability

図1に示す5段階のacceptability評価を実施した．この評価の目的は，文レベルの意味が正しく伝わる文の割合を明らかにすることである．acceptabilityは入力文の意味が正しく伝わらない（たとえば，重要な情報が一つでも欠ける）とFになる。この評価は，adequacyと比べて，より実用に近い評価を目指している．例えば，システムへの要求水準が「入力文の意味が分かればよい」であれば，C以上の評価となった訳文が有用であり，システムへの要求水準が「入力文の意味が分かり，かつ文法的に正しい」であれば，A以上の評価となった訳文が有用である．このように，要求水準に応じた訳質

[*1] http://research.nii.ac.jp/ntcir/ntcir-10/
[*2] http://www.statmt.org/moses/

の文の割合を明らかにすることができる．adequacy 評価の結果からは，このような文の割合は分からない．

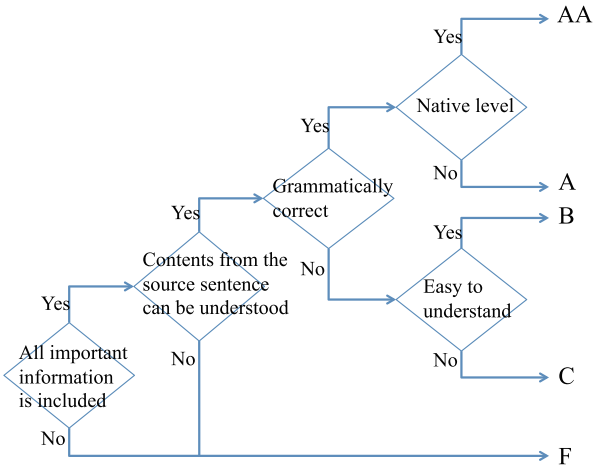

図 1　Acceptability

4　翻訳システムの評価結果

紙面の都合上，評価結果の要点のみ報告する．詳細な報告 [4] および各グループの成果報告はオンラインで入手可能である[*3]．図 2〜4 に adequacy 評価結果，図 5〜7 に acceptability 評価結果を示す．図中のシステム名は，グループ ID（またはシステム ID）とプライオリティ番号の組で表示されている．順位は，adequacy は平均値，acceptability は一対比較に基づいている．機械翻訳における主な課題は訳語選択と語順推定である．SMT は訳語選択の性能は高いため，SMT の現状の大きな課題は語順推定である．以下，各言語対毎に結果を分析する．

4.1　日英翻訳

図 2 および図 5 に評価結果を示す．JAPIO-1 と RBMTx-1 は RBMT，EIWA-1 は RBMT と SMT の HYBRID，KYOTO-1 は EBMT，それ以外は SMT である．RBMT の訳質が SMT より高いことが分かる．acceptability で C 以上の割合は，RBMT のトップのシステムが 6 割程度であったのに対し，SMT のトップのシステムは 2.5 割程度であった．日英翻訳は，言語間の語順が大きく異なる（日本語の語順は SOV，英語の語順は SVO）ため，語順の推定が難しい．それが SMT で高い翻訳品質を達成できなかった主な原因と思われる．

4.2　英日翻訳

図 3 および図 6 に評価結果を示す．RBMTx-1 と JAPIO-1 は RBMT，KYOTO-1 は EBMT，それ以外は SMT である．トップの SMT（NTT-UT-1）の訳質がトップの RBMT と同等以上であることが分かる．英日の特許翻訳で SMT がトップレベルの RBMT と同等以上の訳質を達成したことが明らかになったのはこの評価が初めてである．英日翻訳は，日英翻訳と同様に語順が大きく異なるため，SMT での語順の推定が難しい．そのため，過去の NTCIR-7 では，RBMT の人手評価が SMT より高かった．今回，NTT-UT グループが翻訳の前処理で英語の語順を日本語の語順に入れ替える手法を用いる [8] ことによって，語順の推定精度を大幅に向上させることに成功した．トップの SMT と RBMT のシステムはテスト文の 6 割程度で acceptability が C 以上という結果を得た．

4.3　中英翻訳

図 4 および図 7 に評価結果を示す．RBMTx-1 は RBMT，EIWA-1 は RBMT と SMT の HYBRID，KYOTO-1 は EBMT，それ以外は SMT である．SMT の訳質が RBMT より高いことが分かる．中英翻訳は，日英翻訳に比べて言語間の語順が似ている（どちらも語順が SVO の言語）ことと，RBMT の性能が低かったために，SMT が RBMT よりも良い結果になったと思われる．トップの BBN グループは，特許翻訳の品質向上のために，低頻度な数値表現の汎化，中国語単語分割の最適化，言語モデルの適応，特徴量の追加，英語依存構造の利用などを行い [6]，テスト文の 8 割程度で acceptability が C 以上という結果を得た．

5　自動評価の評価結果

訳質の評価には自動評価が重要である．この重要な役割を担う自動評価の信頼性を人手評価結果に基づいて評価した．自動評価スコアとして BLEU[7]，NIST[1]，RIBES[5] を用いた．詳細なスコア計算方法は，特許機械翻訳タスクの Web ページ[*4]に示されている．テストデータ 2,000 文の訳から計算した各システムの自動評価スコアと各システムの adequacy 平均値とを比較した．人手評価と標準化した自動評価値の散布図を図 8 〜10 に示す．横軸が adequacy 平均値，縦軸が標準化した自動評価値である．表 1 にスピアマン順位相関係数とピアソン積率相関係数，表 2 に RBMT を除いた日英と英日の各係数を示す．

図 8, 9 の緑枠部分は RBMT の結果である．日英・英日では，RBMT の自動評価は人手評価との相関が低い．日英では RBMT を除いても図 8 からいずれの自動評価も中英に比べると人手評価との相関は高くない．英日では RBMT を除くと図 9 から RIBES と人手評価との相関が高いと言える．中英では図 10 からすべての自動評価と人手評価との相関が高いと言える．

6　まとめと今後の予定

NTCIR-9 特許機械翻訳タスクでの分析により，特許翻訳における最新の機械翻訳システムおよび自動評価の評価結果を示した．NTCIR-10 では，より実用的な評価および実際の特許文の傾向を十分に反映した評価の実施を検討中である．

[*3] http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/　[*4] http://ntcir.nii.ac.jp/PatentMT/

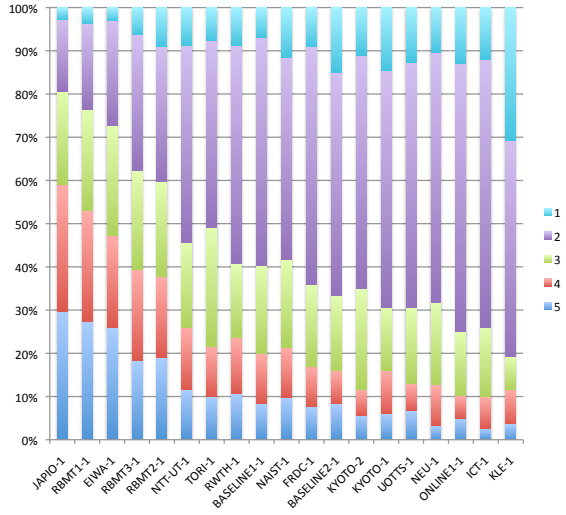

図 2　日英 adequacy

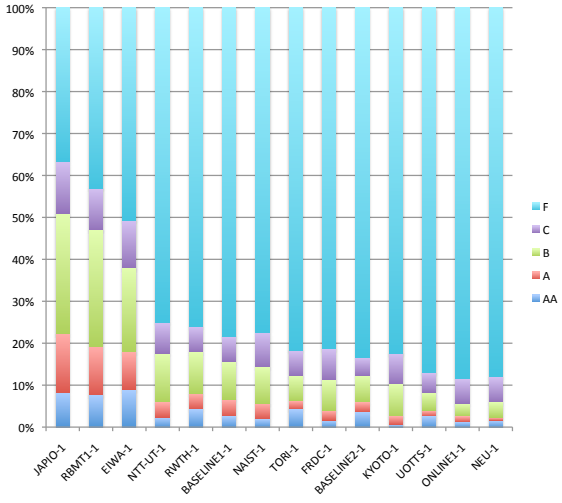

図 5　日英 acceptability

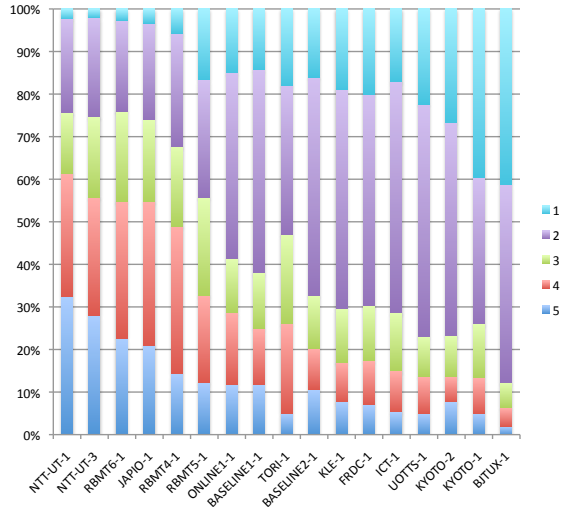

図 3　英日 adequacy

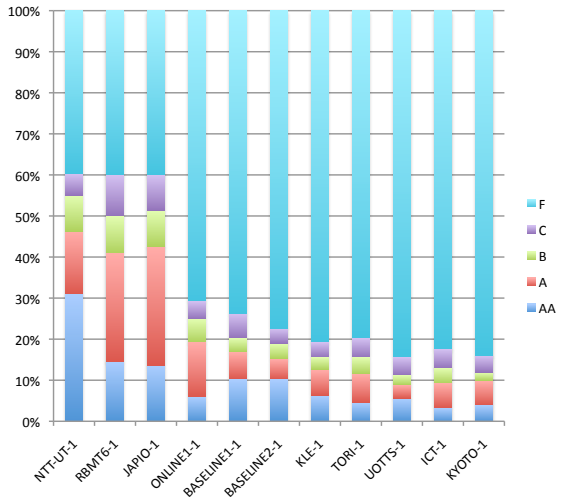

図 6　英日 acceptability

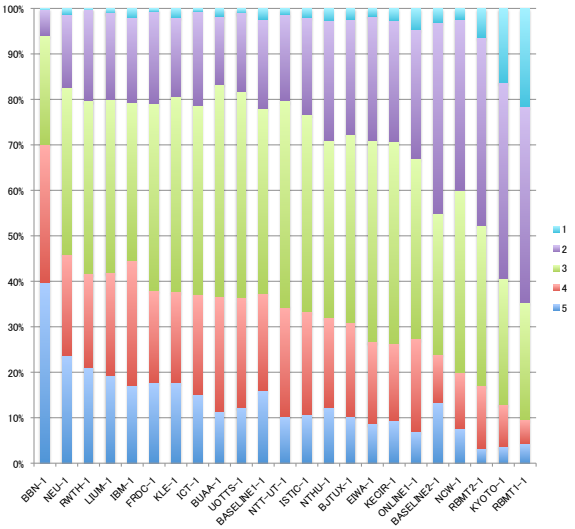

図 4　中英 adequacy

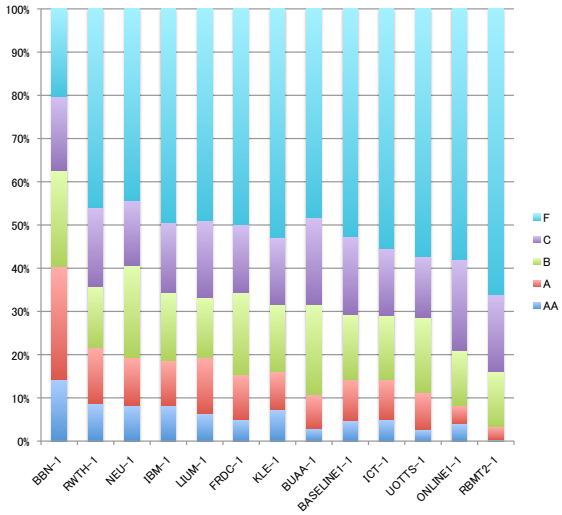

図 7　中英 acceptability

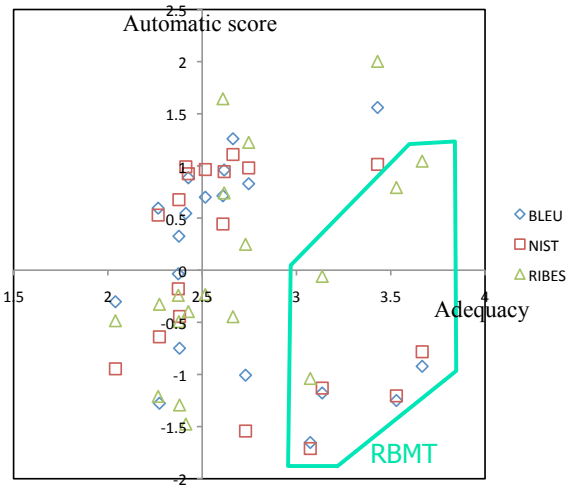

図 8　日英 Adequacy と自動評価との相関

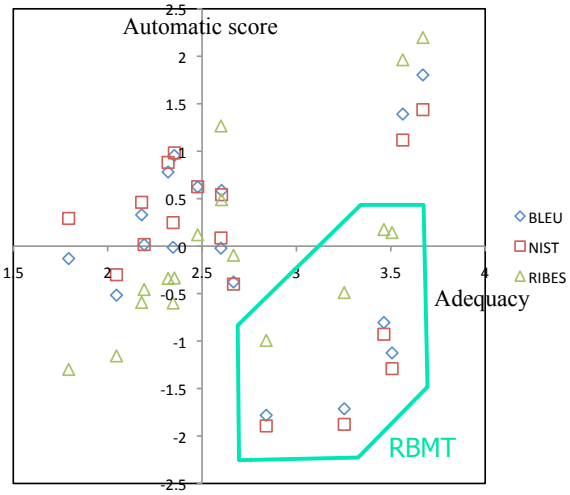

図 9　英日 Adequacy と自動評価との相関

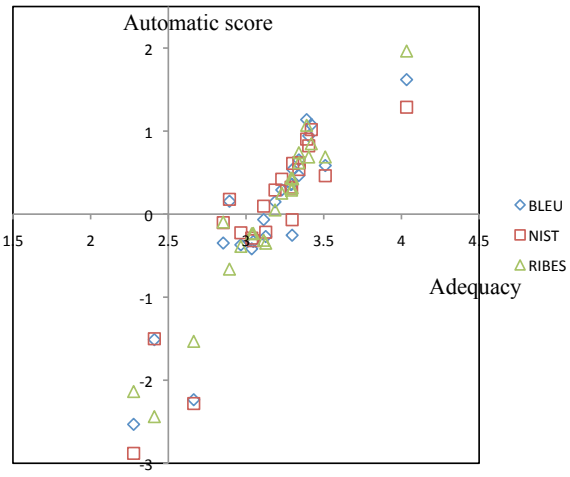

図 10　中英 Adequacy と自動評価との相関

表 1　Adequacy と自動評価の相関係数

		Spearman	Pearson
日英	BLEU	-0.042	-0.241
	NIST	-0.114	-0.286
	RIBES	0.632	0.579
英日	BLEU	-0.029	-0.032
	NIST	-0.074	-0.209
	RIBES	0.716	0.683
中英	BLEU	0.931	0.915
	NIST	0.911	0.891
	RIBES	0.949	0.967

表 2　Adequacy と自動評価の相関係数（RBMT を除く）

		Spearman	Pearson
日英	BLEU	0.618	0.525
	NIST	0.543	0.362
	RIBES	0.679	0.741
英日	BLEU	0.511	0.753
	NIST	0.412	0.603
	RIBES	0.929	0.943

参考文献

[1] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of HLT*, pp. 138–145, 2002.

[2] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. Overview of the Patent Translation Task at the NTCIR-7 Workshop. In *Proceedings of NTCIR-7*, 2008.

[3] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Terumasa Ehara, Hiroshi Echizen-ya, and Sayori Shimohata. Overview of the Patent Translation Task at the NTCIR-8 Workshop. In *Proceedings of NTCIR-8*, 2010.

[4] Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K. Tsou. Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop. In *Proceedings of NTCIR-9*, pp. 559–578, 2011.

[5] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic Evaluation of Translation Quality for Distant Language Pairs. In *Proceedings of EMNLP*, pp. 944–952, 2010.

[6] Jeff Ma and Spyros Matsoukas. BBN's Systems for the Chinese-English Sub-task of NTCIR-9 Patent MT Evaluation. In *Proceedings of NTCIR-9*, 2011.

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*, pp. 311–318, 2002.

[8] Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, Masaaki Nagata, Xianchao Wu, Takuya Matsuzaki, and Jun'Ichi Tsujii. NTT-UT Statistical Machine Translation in NTCIR-9 PatentMT. In *Proceedings of NTCIR-9*, 2011.