特許翻訳における機械翻訳の評価
—NTCIR-10 特許機械翻訳タスクの評価概要—

Evaluation of Machine Translation Systems for Patent Translation

独立行政法人情報通信研究機構　ユニバーサルコミュニケーション研究所多言語翻訳研究室専門研究員　**後藤　功雄**

PROFILE: 1997 年早稲田大学大学院修士課程了。同年 NHK 入局。2008 年より情報通信研究機構に出向。2013 年 9 月 NHK に復帰。自然言語処理の研究に従事。

1 はじめに

筆者らは特許機械翻訳の評価を行っている。これまでに評価型ワークショップ NTCIR[1]-9 および NTCIR-10 にて特許機械翻訳タスク [1,2] を主催し、特許翻訳における機械翻訳システムの評価を実施した。NTCIR-9 でのランダムに選択した文の訳質評価から、現在の最高性能の機械翻訳システムは、特許文（説明文の部分のみでクレーム部分を除く）の翻訳において半数以上の文で原文の内容を理解できることが分かった。この結果から機械翻訳が実用上においても有用である可能性が高いと考え、特許翻訳が必要とされる状況において機械翻訳を利用した場合にどれだけ有用であるかという観点における評価の 1 つとして、NTCIR-10 特許機械翻訳タスクで特許審査での有用性に基づく評価（特許審査評価）を文の訳質評価に加えて実施した。

本稿では、NTCIR-10 特許機械翻訳タスクの概要、評価手法、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は、研究基盤を整備し、複数の翻訳手法に対して同じテストデータを用いた評価を実施して手法の有効性を明らかにすることで、特許翻訳技術の研究・開発を推進することを目的としている。以下にタスク実施の流れを示す。

（1）主催者が特許翻訳用の訓練データとテストデータを用意

（2）参加者が各自のシステムでテストデータを機械翻訳

（3）主催者が翻訳結果を評価

（4）参加者が研究成果をワークショップで発表

この活動を通して構築したデータ（訓練、テスト、評価結果、翻訳結果）は研究利用できるように管理される。NTCIR ワークショップは 1 年半単位で開催されている。第 3 回の NTCIR-3 から特許を対象としたタスクが実施され、第 7 回の NTCIR-7 から特許翻訳のタスクが実施されている。以下、NTCIR-10 特許機械翻訳タスクについて述べる。

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。

実施した評価は、文の訳質評価、特許審査評価、時系列評価およびマルチリンガル評価である。本稿では、これらのうちメインの評価である文の訳質評価と特許審査評価について述べる。

データは次の通りである。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 億文以上を参加者に提供した。開発データには 2,000 文対を用いた。これらのデータは NTCIR-9 で用いたデータと同

1　NII Testbeds and Community for Information access Research

ーである。文の訳質評価のテストデータには新たに構築した 2,300 文、特許審査評価のテストデータには 29 文書を用いた。対訳コーパス、開発データおよびテストデータは、請求項の文を含まず、発明の詳細な説明部分の文からなる。

参加グループ数は全体で 21、日英翻訳は 13、英日翻訳は 11、中英翻訳は 12 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）、RBMT と SMT または EBMT との組み合わせ（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２種類の SMT（フレーズベース SMT、階層フレーズベース SMT）、３つの商用 RBMT[2,3]、Google 翻訳による翻訳結果を追加した。

3 評価手法

3.1 文の訳質評価の手法

3.1.1 評価方法

評価者が文単位の訳質を評価した。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。これらは、主に翻訳結果を読んで内容を理解するという情報アクセスにおける有用性の評価基準である。これらの評価基準は NTCIR-9 で用いた評価基準と同じである。以下、本評価で用いた評価基準およびテストデータの構築方法について説明する。

3.1.2 評価基準

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、訳語選択が適切かだけでなくフレーズや節の意味が正しいかも考慮して評価した。

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。

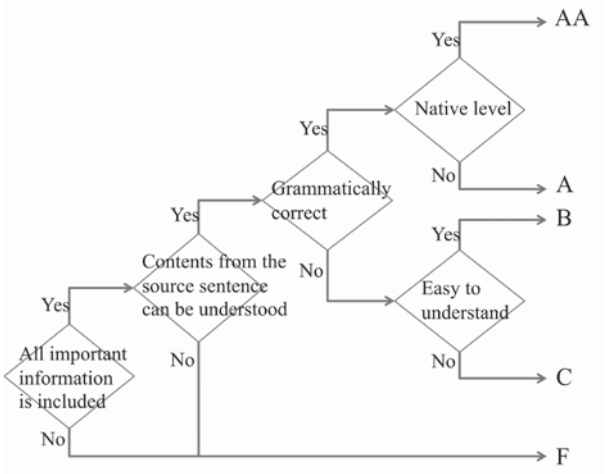

図 1: Acceptability

3.1.3 データの構築

訓練データは 2005 年以前の特許から構築しているため、テストデータは 2006 ～ 2007 年の特許から構築した。我々はテストデータを２つの方法で構築した。１つ目の方法（方法１）は、自動的に抽出した対訳文対からランダムに文対を選択し、さらにその中から人手で正しい対訳の文対を選択するというものである。２つ目の方法（方法２）は、原言語の特許文書からランダムに文を選択し、それらの参照訳は新たに翻訳して構築するというものである。2,300 文のテストデータのうち 2,000 文は方法１で構築し、300 文は方法２で構築した。方法１は、対訳文対を特許文書から自動的に抽出する際に実際の特許文の傾向が完全には反映されなくなる可能性がある。それに対して方法２は、前者の方法

2　３つのうち人手で評価したシステムは NTCIR-9 で最も評価が高かったシステムのみである。

3　中英翻訳では、NTCIR-9 の評価で RBMT の訳質が SMT よりも低かったため、RBMT はベースラインシステムとして利用しなかった。

に比べて実際の特許文の傾向を反映できると考えられる。人手による評価は方法2で作成した300文に対して実施した。この構築方法はNTCIR-9からの改善点である。なお、2,300文のテストデータは自動評価値を計算する際に利用した。

3.2 特許審査評価の手法

特許審査での有用性に基づく評価手法について説明する。特許審査官は、審査において既存の特許を調査して同じ技術が存在すれば、その既存の特許を引用して審査対象の特許を拒絶する。そのため、既存の特許に記載された技術的内容である事実を認定する必要がある。既存の特許が外国語で書かれていてその言語が分からない場合には、翻訳する必要がある。この翻訳を機械翻訳で行った場合に、翻訳結果から引用文書の特許を認定するために有用な事実をどれだけ認定できるかに基づいて、機械翻訳の有用性を評価する。本評価は日本知的財産翻訳協会（NIPTA）の協力により実施した。

3.2.1 全体の評価の流れ

ここで、全体の評価の流れについて説明する。

（1）準備　評価で利用するデータを構築する。まず審査の結論が拒絶の審決（審査の最終決定）とその審査で引用された特許を取得する。そして、審査において「引用文書から審査官が認定した事実」の根拠となる文を引用文書から抽出する。抽出した文をテストデータとする。なぜなら、この抽出した文は、「審査官が認定した事実」を表しているためである。そして、このテストデータが正しく翻訳されれば、翻訳結果から「審査官が認定した事実」を認定することができるためである。

（2）翻訳　テストデータを機械翻訳する。

（3）評価　翻訳結果から、「引用文書から審査官が認定した事実」をどれだけ認定できて、審査に有用であるかについて評価する。

3.2.2 評価方法

特許審査評価は日英翻訳と中英翻訳に対して行った。特許庁で審査官の経験があり、英語の能力が高い2人の評価者が評価した。1システムあたり各評価者が20文書、合計40文書（文書の重複を含む）の評価を行なった。評価したシステムは、adequacyの結果が上位のシステムから手法の多様性を考慮して日英、中英それぞれ3システムを選択した。

3.2.3 評価基準

特許審査評価で用いた評価基準を表1に示す。有用性の評価は、過去の審査で審査官が引用文書の特許から認定した事実を、引用文書を機械翻訳した結果からどれだけ認定できるかに基づいて行った。評価は引用文書単位で行なった。

3.2.4 データの構築

この評価に必要なデータを、審決を用いて以下のようにして構築した。

（1）結論が不成立（拒絶）の審決を取得する。

（2）審決中に記載されている「引用文書から審査官が認定した事実の説明」を抽出する。

（3）審査官が認定した事実を構成要素単位に分けて、それぞれの内容の根拠となる文を引用文書から抽出する。抽出した文を機械翻訳で翻訳するテストデータとする。

表2に、審査官が認定した事実と引用文書から抽出した文の例を示す。表2の左端の列は、審決中に記載されている審査官が引用文書から認定した事実である。表2の中央の列は、審査官が認定した事実を構成要素単位に分けたものである。表2の右端の列は、中央の列の事実を認定する根拠となった引用文書中の文を抽出したものであり、この文が翻訳するテストデータである。29文書のテストデータを構築した。中英翻訳のテストデータは、日英翻訳のテストデータを日中翻訳して構築した。

4 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[2]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[4]。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する。

4 http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/

表 1：特許審査評価の評価基準

Ⅵ	引用発明を認定するために有用な事実が全て認定できて，翻訳結果のみで審査可能
Ⅴ	引用発明を認定するために有用な事実が半分以上認定できて，審査に有用
Ⅳ	引用発明を認定するために有用な事実が 1 つ以上認定できて，審査に有用
Ⅲ	Ⅳに至らないが，部分的に事実が認定できて，その文献が審査で無視できないことが分かる
Ⅱ	一部の事実が認定できたが，審査に有用とはいえない
Ⅰ	全く事実が認定できず，審査の役に立たない

（引用文書単位の評価）

表 2：審査官が認定した事実と引用文書から抽出した文の例

審査官が認定した事実	構成要素単位に分けた審査官が認定した事実	引用文書から抽出した文（テストデータ）
これらの記載事項によると、引用例には、「内部において、先端側に良熱伝導金属部 43 が入り込んでいる中心電極 4 と、中心電極 4 の先端部に溶接されている貴金属チップ 45 と、中心電極 4 を電極先端部 41 が碍子先端部 31 から突出するように挿嵌保持する絶縁碍子 3 と、絶縁碍子 3 を挿嵌保持する取付金具 2 と、中心電極 4 の電極先端部 41 との間に火花放電ギャップ G を形成する接地電極 11 とを備えたスパークプラグにおいて、中心電極 4 の直径は、1.2 ～ 2.2mm としたスパークプラグ。」の発明が記載されていると認められる。	内部において、先端側に良熱伝導金属部 43 が入り込んでいる中心電極 4	また、図 3 に示すごとく、中心電極 4 の内部においては、上記露出開始部 431 よりも先端側にも良熱伝導金属部 43 が入り込んでいる。
	中心電極 4 の先端部に溶接されている貴金属チップ 45	また、中心電極 4 の先端部には、貴金属チップ 45 が溶接されている。
	中心電極 4 を電極先端部 41 が碍子先端部 31 から突出するように挿嵌保持する絶縁碍子 3	上記中心電極 4 は、電極先端部 41 が碍子先端部 31 から突出するように絶縁碍子 3 に挿嵌保持されている。
	絶縁碍子 3 を挿嵌保持する取付金具 2	上記絶縁碍子 3 は、碍子先端部 31 が突出するように取付金具 2 に挿嵌保持される。
	中心電極 4 の電極先端部 41 との間に火花放電ギャップ G を形成する接地電極 11	上記接地電極 11 は、図 2 に示すごとく、電極先端部 41 との間に火花放電ギャップ G を形成する。
	中心電極 4 の直径は、1.2 ～ 2.2mm	また、上記碍子固定部 22 の軸方向位置における中心電極 4 の直径は、例えば、1.2 ～ 2.2mm とすることができる。

4.1　文の訳質評価の結果

評価結果を図 2 から 7 に示す。Adequacy の評価には平均値を用いた。

4.1.1　日英翻訳

図 2 と 3 が日英翻訳の評価結果である。JAPIO-1 と RBMT1-1 は RBMT、EIWA-1 と TORI-1 は 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 が C 以上の割合は、RBMT のトップのシステムが 55% であったのに対し、SMT のトップのシステムは 38% であった。C 以上の割合における RBMT に対する SMT の割合は NTCIR-9 では 39%（0.25/0.633）であったのに対し、NTCIR-10 では 69%（0.38/0.55）であり、割合が高くなっている。これは、RBMT に対して SMT の性能が向上していることを示している。NTCIR-10 でトップの RBMT の性能は NTCIR-9 でトップの RBMT と比べて同等以上と考えられるため、NTCIR-10 でトップの SMT の性能は NTCIR-9 でトップの SMT に対して向上しているといえる。

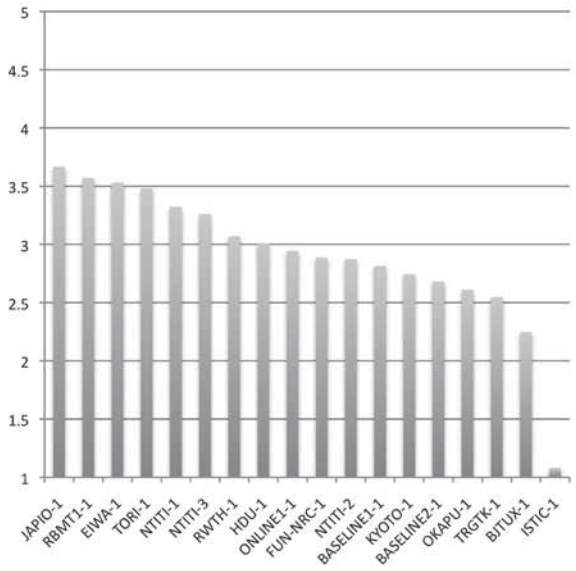

図 2: 日英 adequacy

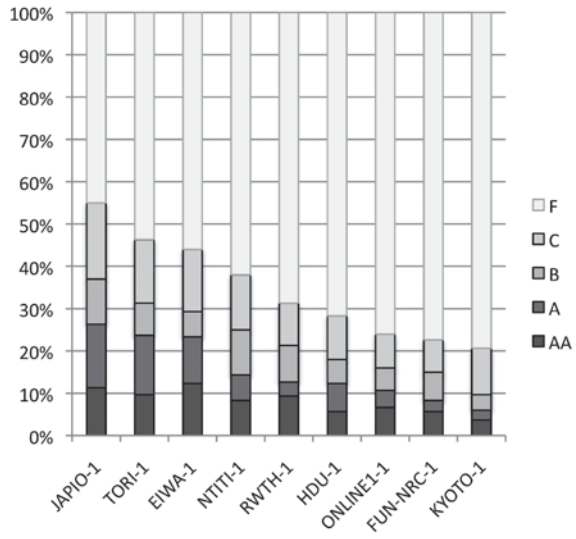

図 3: 日英 acceptability

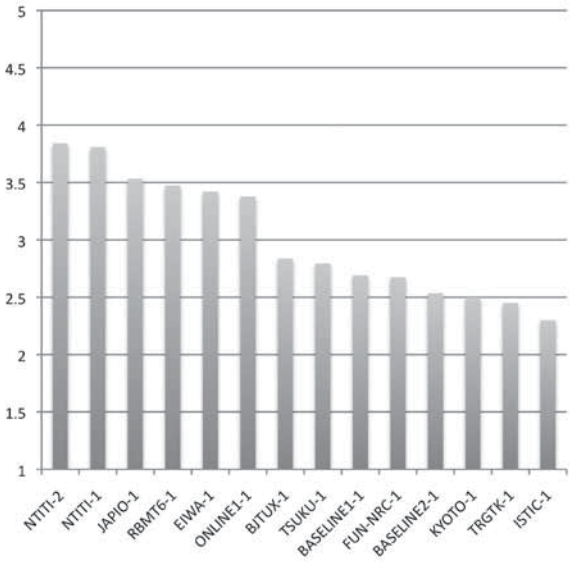

図 4: 英日 adequacy

4.1.2　英日翻訳

図 4 と 5 が英日翻訳の評価結果である。JAPIO-1 と RBMT6-1 は RBMT、KYOTO-1 は EBMT、それ以外は SMT である。トップの SMT（NTITI-2 および NTITI-1）の訳質がトップの RBMT よりも高いことが分かる。NTCIR-9 では、英日の特許翻訳で SMT がトップレベルの RBMT と同程度の訳質であったため、NTCIR-9 と比べてトップの SMT の性能が向上していることが分かる。NTITI のシステムは、NTCIR-9 で NTT-UT システムが利用した「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、さらに特許文の構文解析精度を向上させることによって、訳質を向上させた。トップの SMT のシステムはテスト文の 70%、トップの RBMT は 58% で acceptability が C 以上という結果を得た。

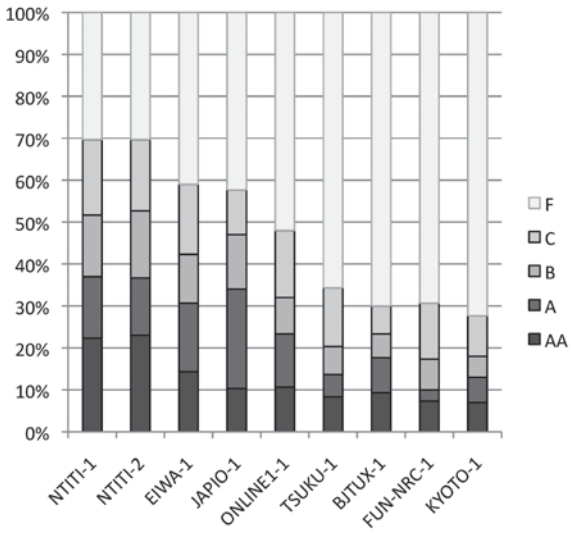

図 5: 英日 acceptability

4.1.3　中英翻訳

図 6 と 7 が中英翻訳の評価結果である。RWSYS-1 と EIWA-1 は HYBRID、BJTUX-2 は EBMT、それ以外は SMT である。トップの BBN-1 システムは、NTCIR-9 での BBN-1 システムに対して、言語モデルや翻訳モデルなどをさらに改良し、テスト文の 67% で acceptability が C 以上という結果を得た。

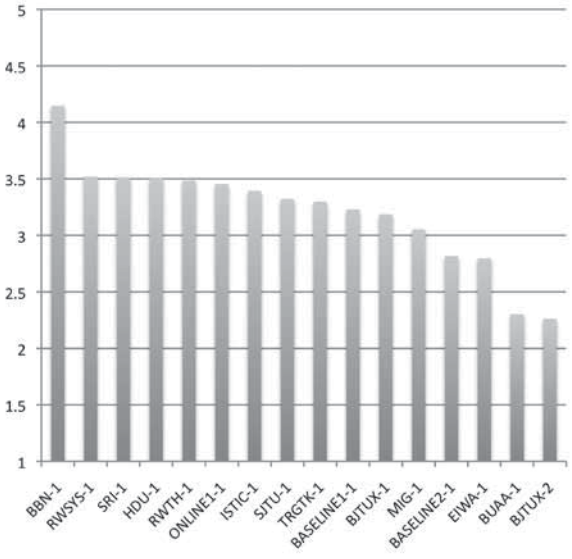

図 6: 中英 adequacy

4.2　特許審査評価の結果

日英翻訳の評価結果を図 8 に、中英翻訳の評価結果を図 9 に示す。図 8 より、JAPIO-1 システムは 6 割以上の文書で評価 VI を獲得し、全文書が評価 V 以上で

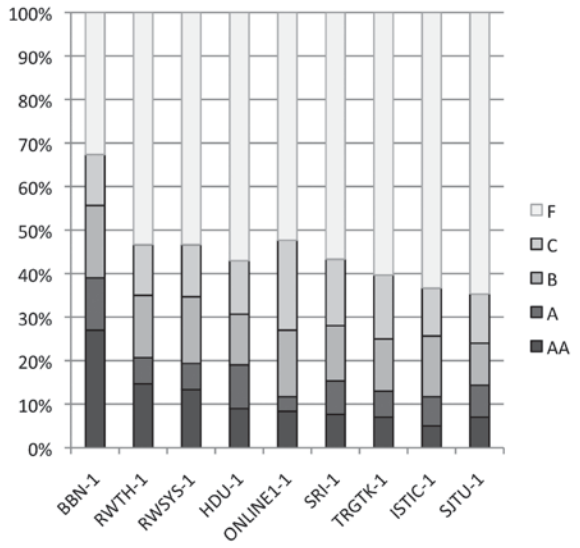

図 7: 中英 acceptability

あった。これより、トップの日英翻訳システムの性能は特許審査で有用なレベルであることが分かった。SMT の NTITI-1 システムも評価 V 以上が 6 割以上あり、ある程度有用であることが分かった。また、図 9 より、BBN-1 システムは 2 割の文書で評価 VI を獲得し、9 割弱の文書で評価 V 以上であった。これより、トップの中英翻訳システムの性能も特許審査で有用なレベルであることが分かった。

5 まとめ

NTCIR-10 特許機械翻訳タスクでの評価について述

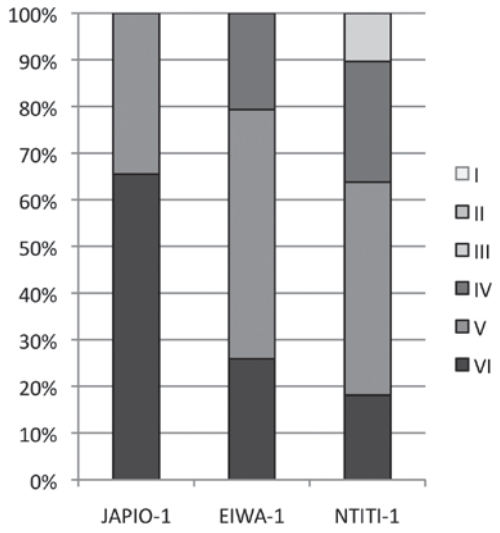

図 8: 日英 特許審査評価の結果

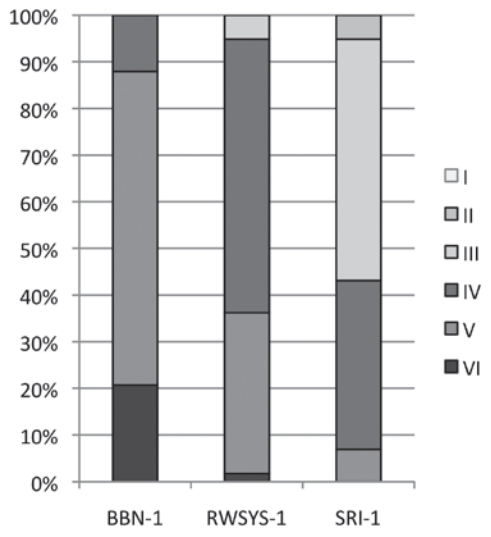

図 9: 中英 特許審査評価の結果

べた。この評価で実施した特許審査評価により、特許審査において、現在のトップレベルの機械翻訳システムが日英・中英翻訳で有用であることが明らかになった。また、英日翻訳ではトップレベルの SMT が RBMT よりも特許翻訳で高い訳質を達成したことが分かった。日英翻訳では、RBMT がトップレベルの SMT より依然として訳質が高かったが、SMT の性能が NTCIR-9 の時よりも向上していることが確認された。

謝辞

本評価に協力していただいた評価者およびデータ作成者に感謝します。また、特許審査評価のテストセット作成および評価に協力していただいた日本知的財産翻訳協会（NIPTA）石井理事長他に感謝します。

参考文献

[1] 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.

[2] Isao Goto, Ka Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K. Tsou. Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop. In Proceedings of NTCIR-10, pp. 260–286, 2013.