大規模な共通基盤による機械翻訳システムの比較評価 NTCIR特許翻訳タスク最新事情

藤井 敦 † 内山 将夫 ‡ 山本 幹雄 † 宇津呂 武仁 †
† 筑波大学 ‡ 情報通信研究機構

1 はじめに

自然言語処理や情報検索などの言語情報処理に関する研究では,「言葉の意味」,「感情」,「情報要求」といった,厳密な定義が困難な概念を研究の対象としている．しかし，科学や工学における研究分野として言語情報処理を位置付けるためには，問題の定式化や評価において，学問として要求される水準を満たす必要がある．すなわち，学術研究としての実証性，客観性，再現性が求められている．データマイニングの国際会議 KDD2009 では，手法の新規性などに加えて，再現性（repeatability）が採択の要件である[1]．こうした動向は言語情報処理の分野にも影響する可能性がある．

そこで，複数の研究者が共有できる評価基盤として，大規模かつ再利用可能なテストコレクションが重要である．テストコレクションを組織的に構築するために，評価ワークショップという活動形態が存在する．評価ワークショップでは，あるタスクについて，複数の研究グループが共通のデータを使用することで透明性が高いシステムの比較を行う．その過程を通して，対象のタスクに適したテストコレクションと評価方法を開発する．

筆者らは，NTCIR ワークショップにおいて，特許情報を対象としたテストコレクションの構築研究を行っている [9]．本稿執筆当時に最新の第 7 回 NTCIR ワークショップ（NTCIR-7）では，機械翻訳のタスクを対象としたテストコレクションを構築した．特許翻訳のタスクには，機械翻訳の学術研究や開発が促進されるという効果がある．また，海外に出願された特許を検索したり，海外に出願するために日本語の特許を翻訳するための基盤技術になるため，産業上の効果がある．

本稿は，特許翻訳タスクによって得られた知見と今後の活動計画について記述する．

2 NTCIR 機械翻訳タスクの概要

特許翻訳タスクでは，複数の参加グループがシステムの訓練と評価に共通のデータを使用することで，手法やシステムの体系的な比較評価を容易にした．

近年，統計的な機械翻訳（SMT）が発展している．SMT が有効に機能するためには，訓練データとして，直訳に近い大量の対訳文が必要である．日本語について，この条件を満たす訓練データは，IWSLT[2]の旅行会話文などに限られていた．

折しも，筆者らは特許情報から上記の条件を満たす訓練データを構築できる点に着目し，特許翻訳タスクを実行するに至った．日本に出願される発明のうち，一定の件数は外国語に翻訳されて海外にも出願される．同じ発明に対して出願された特許の集合は「パテントファミリー」と呼ばれる．

多くの場合，パテントファミリーは優先権主張番号という識別番号で特定することが可能である．パテントファミリーにおいて,「背景（Background of the Invention）」と「実施例（Detailed Description of the Preferred Embodiments）」の項目は文単位で直訳される傾向にある．そこで，文対応付け手法 [7] をこれらの項目に適用し，日本語と英語の文対応データを抽出した．

具体的には，日本公開特許公報と米国特許公報を用いて，1993–2000 年発行分と 2001–2002 年発行分から独立に文対応データを抽出し，それぞれを「訓練データ」と「テストデータ」として使用した．すなわち，発行済みの特許公報を用いて MT システムを学習し，新規発行の特許公報を翻訳する状況を想定している．訓練データは，日本語と英語の約 180 万文対であり，日英の対訳文データとしては大規模である．3000 文を抽出して調査した結果，約 90%の文対応が正しい翻訳対であった．また，テストデータとして，正しい翻訳対と認められた文対応だけを用いた．テストデータは 1381 文対ある．

訓練データは，特許翻訳タスクの参加グループに事前に配布され，各グループは自らのシステムを訓練することが許された．なお，参加グループは，SMT だけでなく，規則に基づく MT（RBMT）や事例に基づく MT（EBMT）といった他の手法を使用してもよい．

評価手法として,「訳質による直接的な評価（Intrinsic 評価）」と「応用による間接的な評価（Extrinsic 評価）」を行った．3 節と 4 節で各評価手法について説明する．

[1] http://www.kdd.org/kdd2009/

[2] http://www.slc.atr.jp/IWSLT2008/

3 Intrinsic 評価

機械翻訳の訳質を評価するための唯一絶対的な基準は確立されていない．特許翻訳タスクでは，近年よく用いられている複数の評価尺度を用いた．さらに，異なる評価尺度によって参加グループの比較評価がどのように変化するのかを調査した．

具体的には，人手判定による評価と自動的な評価尺度である BLEU [6] を個別に使用した．人手判定による評価では，3 名の判定者が独立に作業を行い,「翻訳の適切さ（Adequacy）」と「目的言語における流暢さ（Fluency）」に対する 5 段階評価を行った．ただし，作業コストを制限する理由から，テストデータから無作為に 100 文対を抽出し，人手判定の対象とした．

2 節で説明したように，2001–2002 年発行分の特許公報から日英の 1381 文対をテストデータとして抽出した．日本語と英語の一方で書かれた文を MT システムへの入力として使い，他方の言語で書かれた文を参照訳として BLEU を計算した．以下，テスト文対の実例を示す．

- 図５は回転羽根２を駆動するモータの構成例を示す図である。

  FIG. 5 is a diagram showing a structural example of a motor for driving the rotating blade 2.

- さらに、心線ワイヤ５１の先端部がラミネートフィルム５９により挟まれて保持され、その変形、ピッチの狂いを防止する。

  Moreover, the front ends of the core wires 51 are sandwiched with laminated films 59 to prevent deformation of the core wires 51 for the purpose of maintaining their relative positions intact.

- この絶縁ハウジング１０の外面に取り付けられるシールドカバー３０を図６に示している。

  FIG. 6 shows the shield cover 30, which is to be mounted on the insulative housing 10.

しかし，1 つのテスト文には複数の正しい翻訳が存在する．そこで，テスト文あたりの参照訳数を増やすために，特許の翻訳家 2 名に翻訳作業を依頼した．2 名の翻訳家は同一のテスト文集合を独立に翻訳した．作業コストを制限する理由から，1381 文対から無作為に 300 文対を選び，日本語のテスト文を英語に翻訳してもらった．

特許の翻訳家は，通常の業務において辞書や MT システムなどの翻訳ツールを使用することがある．しかし，特定の MT システムに対して有利な参照訳を作らないようにするため，翻訳家は今回の作業では MT システムを使用しなかった．

4 Extrinsic 評価

Extrinsic 評価では，MT を言語横断情報検索（Cross-Lingual Information Retrieval: CLIR）に応用し，検索精度によって MT を間接的に評価した．検索精度として MAP（Mean Average Precision）を使用した．

CLIR の検索精度を評価するために，NTCIR-5 の特許翻訳タスク [2] で構築したテストコレクションを使用した．ある発明が特許として成立し，その権利が消滅する過程では様々な調査が行われる．調査の目的に応じて，性質の異なる特許検索が必要になる．NTCIR-5 のテストコレクションは，他者の出願を無効にするための「無効資料調査」の検索精度を評価することを目的としている．すなわち，1 つの検索課題は，日本公開特許公報中の請求項 1 件であり，さらに特許の翻訳家によって英語に翻訳された．検索対象の文書データは，1993–2002 年発行分の日本公開特許公報である．

そこで，日本語による単言語検索の精度と英日 CLIR の精度を評価することができる．さらに，元の日本語請求項を参照訳として使用することで，請求項の英日翻訳に対する BLEU を計算した．

検索課題は 1189 件あるものの，翻訳にかかる計算時間を考慮して，検索課題の件数を制限した．具体的には，日本語単言語検索の精度が中程度の 124 件を選択した．一般に，請求項は長く複雑な名詞句である．以下，検索課題として使用した請求項の一例を示す．

> A multiphase structure carbon electrode material made of a carbonaceous material having multilayer structure formed by covering a carbonaceous material having high crystallinity with a carbonaceous material having relatively low crystallinity, in which the covered layer is not broken to expose the core material.

参加グループの多くが情報検索システムの構築には馴染みがなかった．そこで，参加グループは英語の検索課題を日本語に翻訳してオーガナイザ（本稿の筆者ら）に提出し，オーガナイザは全グループに対して共通のシステムを用い，検索を代行した．すなわち，Extrinsic 評価におけるグループ間の優劣は MT だけで決まる．

5 評価結果

日英 Intrinsic，英日 Intrinsic，英日 Extrinsic への参加グループ数はそれぞれ 14，12，12 であり，総合すると異なりで 15 グループが参加した．さらに，オーガナイザが Moses を用いた結果を追加した．紙面の都合上，

評価結果の要点のみ報告する．詳細な報告 [3] 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]．

表 1 にテスト文と検索課題の長さを示す．日本語と英語のデータ長は，それぞれ文字と単語で計上した．Intrinsic と Extrinsic に使用した英語のデータ長を比較すると，Extrinsic 評価に使用した検索課題が長いことが分かる．

表 1: テスト文と検索課題の長さ

種別	単位	最短	平均	最長
Intrinsic 日本語	文字	11	60.1	302
Intrinsic 英語	単語	5	29.0	117
Extrinsic 英語	単語	13	115.4	412

まず，日英 Intrinsic 評価の結果について説明する．図 1 は，参加グループごとに BLEU の 95%信頼区間 [4] を示しており，参加グループは BLEU の値で降順に並んでいる．tsbmt と JAPIO は RBMT を使用し，Kyoto-U は EBMT を使用した．残りのグループは全て SMT を使用した．括弧が付いているグループ名は，締切後に提出された結果であり，公式結果ではない．また，左から 2 番目の「Moses *」は，オーガナイザが追加した Moses による結果である．以降の図においても同じ表記を用いる．

図 1 は，1381 の日本語文に対して，パテントファミリーから対訳文として抽出された英文だけを参照訳とした場合の結果である．上位 3 グループの BLEU はそれぞれ 27.20，27.14，27.14 だった．最左の NTT [8] は階層型フレーズ SMT を使用し，Moses と MIT はフレーズ SMT を使用した．図 2 は，300 の日本語文に対して，2 名の翻訳家が作成した参照訳を併用した場合の結果である．パテントファミリーから抽出した対訳文は参照訳として使用しなかった．特許出願時に MT システムが下訳に使われることがあるため，特定の MT システムに有利な評価を避けるためである．

図 1 と図 2 において，参加グループの並び順は同じである．そこで，参照訳のパターンを増やしたことによって，BLEU の大小関係がどのように変動したかを見比べることができる．図 2 では，RBMT を使用した tsbmt と JAPIO に対する BLEU の増加が顕著だった．

図 3 に人手判定の結果を示す．図 3 においても参加グループの並び順は図 1 と同じである．図 3 の縦軸は，3 名の判定者による結果を Adequacy と Fluency ごとに平均し，さらに Adequacy と Fluency を平均した値である．RBMT を使用した tsbmt と JAPIO に対する判定値が高いことが分かる．

BLEU が SMT に対して高い評価を与え，人手判定が

[3]http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/

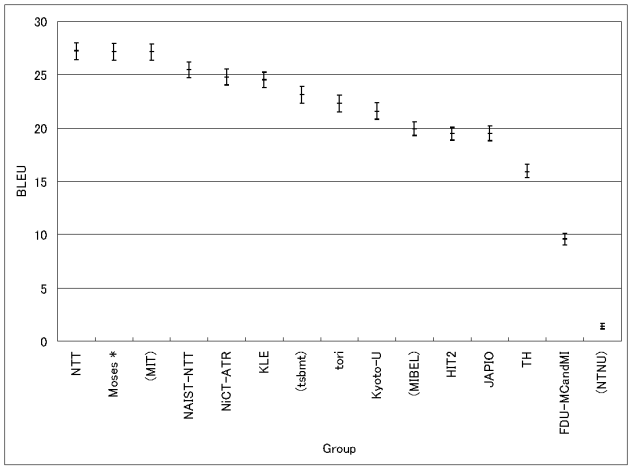

図 1: 日英 Intrinsic: BLEU 95%信頼区間（参照訳× 1）

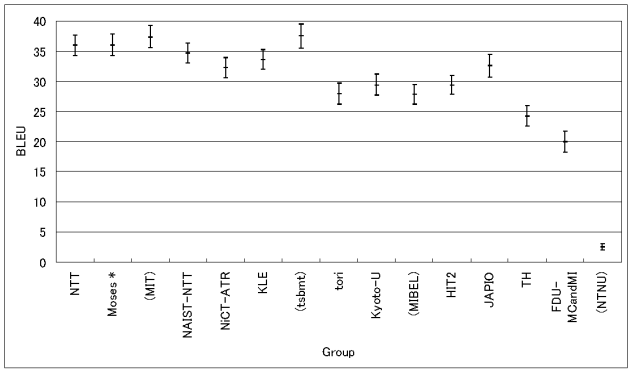

図 2: 日英 Intrinsic: BLEU 95%信頼区間（参照訳× 2）

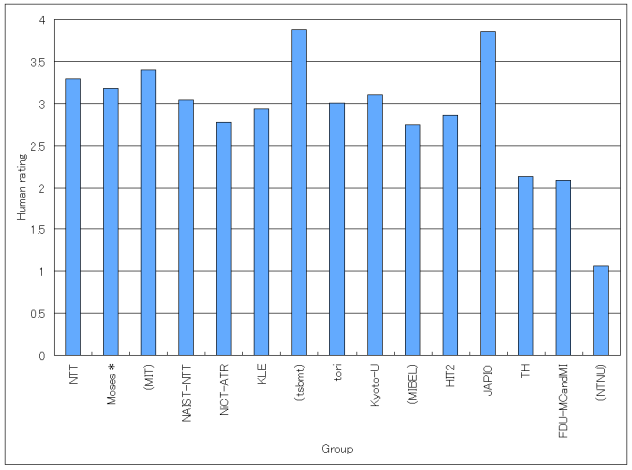

図 3: 日英 Intrinsic: 人手判定による評価

RBMT に対して高い評価を与えることは，欧米言語間の MT 研究でも報告されている [1, 5]．NTCIR-7 特許翻訳タスクによって，日英の特許情報を対象にした MT でも同様の傾向が確認された．

人手判定による評価と BLEU による評価の関係について調べるために，相関係数を計算した．人手判定による評価と図 1 の BLEU に対する相関係数は 0.814 であり，人手判定による評価と図 2 の BLEU に対する相関係数は 0.909 であった．すなわち，参照訳のパターンを増やすことによって，人手判定と BLEU による評価結果は近付いた．しかし，図 2 は 300 のテスト文しか使用していないことと，SMT 以外の手法を用いたグループが少ないこととから，今後もさらに研究が必要である．

次に，英日 Intrinsic 評価の結果について説明する．ここでは，パテントファミリーから抽出された対訳文だけを参照訳として使用した．上位 3 グループは，Moses，HCRL NiCT-ATR であり，BLEU の値はそれぞれ 30.58，29.97，29.15 だった．なお，HCRL は日英 Intrinsic 評価には参加しなかった．

さらに，英日 Intrinsic 評価と英日 Extrinsic 評価の両方に参加したグループの BLEU を比較した．その結果，相関係数は 0.964 という高い値であった．訓練データは「背景」と「実施例」から抽出されており，Extrinsic 評価に用いた「請求項」とは異質であるにもかかわらず，Intrinsic 評価の BLEU が高いグループは Extrinsic 評価の BLEU も高い傾向にあることが分かった．

最後に，英日 Extrinsic 評価の結果について説明する．Extrinsic 評価では，CLIR の検索精度を MAP で評価し，さらに請求項に対する翻訳精度を BLEU で評価した．12 の参加グループについて MAP と BLEU の相関係数は 0.936 という高い値であった．すなわち，BLEU の値が高いほど，CLIR の検索精度も高いことが分かった．

6 まとめと今後の予定

5 節の結果から，異なる評価尺度の関係について分かった点をまとめる．

- BLEU と MAP による評価結果は相関が高かった．
- BLEU と人手判定による評価結果の相関は低かった．BLEU を用いると SMT の評価が高くなり，人手判定では RBMT の評価が高い傾向にあった．しかし，参照訳のパターンを増やすことによって，BLEU と人手判定による結果は近付いた．

2 つ目の点は検討に値する．人手判定とほぼ同じ評価結果を再現することができる自動評価手法が確立されれば，再利用可能なテストコレクションができ，その結果，MT の研究開発を効率的に進めることができる．

そこで，次回の第 8 回 NTCIR ワークショップでは，特許翻訳タスクを継続するとともに,「評価手法を探求するサブタスク」を新たに実践する．ここでは，人手判定に近い評価結果を自動的な手法で再現することが目的である．さらに，対象のデータを拡張することも計画している．具体的には，特許公報の発行年数を増やすことや日英以外の言語も対象とすることを予定している．

謝辞

本研究の一部は，平成 20 年度国立情報学研究所共同研究「特許文書における統計的機械翻訳技術の評価と協調的研究基盤資源の構築」で行った．

参考文献

[1] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 249–256, 2006.

[2] Atsushi Fujii, Makoto Iwayama, and Noriko Kando. Test collections for patent retrieval and patent classification in the fifth NTCIR workshop. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pp. 671–674, 2006.

[3] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. Overview of the patent translation task at the NTCIR-7 workshop. In *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, pp. 389–400, 2008.

[4] Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 388–395, 2004.

[5] Philipp Koehn and Christof Monz. Manual and automatic evaluation of machine translation between European languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pp. 102–121, 2006.

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[7] Masao Utiyama and Hitoshi Isahara. A Japanese-English patent parallel corpus. In *Proceedings of MT Summit XI*, pp. 475–482, 2007.

[8] Taro Watanabe, Hajime Tsukada, and Hideki Isozaki. NTT SMT system 2008 at NTCIR-7. In *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, pp. 420–422, 2008.

[9] 藤井敦, 難波英嗣, 岩山真, 神門典子, 内山将夫, 山本幹雄, 宇津呂武仁, 橋本泰一. 特許情報処理を指向したテストコレクションの構築：情報検索と自然言語処理の融合を目指して．情報処理学会研究報告, 2008-FI-89/2008-NL-183, pp. 31–36, 2008.