平成 24 年度 AAMT/Japio 特許翻訳研究会

報　告　書

機械翻訳及び辞書構築に関する研究
及び
シンポジウム・拡大評価部会報告

平成 25 年 3 月

一般財団法人 日本特許情報機構

目　　次

AAMT／Japio 特許翻訳研究会委員名簿

（敬称略・順不同）

委 員 長	辻井　潤一	マイクロソフトリサーチアジア・東京大学名誉教授・ マンチェスター大学客員教授・AAMT 元会長
副委員長	横山　晶一	山形大学大学院教授
〃	江原　暉将	山梨英和大学教授
委　　員	宮澤　信一郎	秀明大学教授
〃	梶　　博行	静岡大学教授
〃	黒橋　禎夫	京都大学大学院教授
〃	宇津呂　武仁	筑波大学教授
〃	二宮　崇	愛媛大学大学院准教授
〃	越前谷　博	北海学園大学准教授
〃	綱川　隆司	静岡大学助教
〃	範　　暁蓉	東京大学大学院　中川研究室
〃	後藤　功雄	(独)情報通信研究機構
〃	熊野　明	東芝ソリューション(株)
〃	下畑　さより	沖電気工業(株)
〃	潮田　明	(株)富士通研究所
〃	三浦　貢	日本電気(株)
〃	須藤　克仁	NTT コミュニケーション科学基礎研究所
事 務 局	村上　嘉陽	AAMT/Japio 特許翻訳研究会東京事務局・(株)ナビックス
〃	河田　容英	〃　　〃　　・(株)ログワークス
〃	高田　佳代子	〃　　〃
オブザーバー	中川　裕志	東京大学大学院教授
〃	安藤　進	元多摩美術大学講師
〃	呉　先超	バイドゥ(株)
〃	守屋　敏道	(財)日本特許情報機構
〃	松田　成正	〃
〃	大塩　只明	〃
〃	塙　　金治	〃
〃	三橋　朋晴	〃
〃	柿田　剛史	〃
〃	土屋　雅史	〃
〃	星山　直人	〃
〃	王　　向莉	〃

１．　平成 24 年度 AAMT/Japio 特許翻訳研究会・活動履歴

平成 24（2012）年 4 月 20 日
第 1 回 AAMT/Japio 特許翻訳研究会（於キャンパス・イノベーションセンター東京）

平成 24（2012）年 6 月 1 日
第 2 回 AAMT/Japio 特許翻訳研究会（於キャンパス・イノベーションセンター東京）

平成 24（2012）年 7 月 6 日
第 3 回 AAMT/Japio 特許翻訳研究会（於キャンパス・イノベーションセンター東京）

平成 24（2012）年 9 月 7 日
特許文書の機械翻訳結果評価方法検討会（於東京大学工学部 11 号館 1 階講堂）

平成 24（2012）年 10 月 12 日
第 4 回 AAMT/Japio 特許翻訳研究会（於キャンパス・イノベーションセンター東京）

平成 24（2012）年 11 月 30 日
第 2 回特許情報シンポジウム（於京都大学東京オフィス）

平成 24（2012）年 12 月 7 日
第 5 回 AAMT/Japio 特許翻訳研究会（於中央大学駿河台記念館）

平成 25（2013）年 1 月 11 日
AAMT/Japio 特許翻訳研究会・拡大評価部会（於キャンパス・イノベーションセンター東京）

平成 25（2013）年 3 月 8 日
第 6 回 AAMT/Japio 特許翻訳研究会（於キャンパス・イノベーションセンター東京）

平成 25（2013）年 3 月 29 日
『平成 24 年度 AAMT/Japio 特許翻訳研究会報告書　　機械翻訳及び辞書構築に関する研究及びシンポジウム・拡大評価部会報告』完成

以　上

２．１　日英パテントファミリーにおける対訳文対非抽出部分を利用した専門用語訳語推定

筑波大学大学院システム情報工学研究科
豊田　樹生，牧田　健作，高橋　佑介
龍　梓，董　麗娟，宇津呂　武仁，山本　幹雄

2.1.1　はじめに

特許文書の翻訳は，他国への特許申請や特許文書の言語横断検索などといったサービスにおいて不可欠である．特許文書翻訳の過程において，専門用語の対訳辞書は重要な情報源であり，これまでに，対訳特許文書を情報源として，専門用語対訳対を自動獲得する手法の研究が行われてきた．文献[4]では，NTCIR-7 特許翻訳タスク[1] において配布された日英 180 万件の対訳特許文を用いて，対訳特許文からの専門用語対訳対獲得を行った．この研究では，句に基づく統計的機械翻訳モデル[2]を用いることにより，対訳特許文から学習されたフレーズテーブル，要素合成法，Support Vector Machines (SVMs) [8] を用いることによって，専門用語対訳対獲得を行った．また，文献[3]においては，文献[4]の専門用語訳語推定タスクの後段のタスクとして，同義対訳専門用語の同定と収集を行っている．

ここで，上述の日英 180 万件の対訳特許文は，文献[7]の手法により，日米パテントファミリーの対応特許文書中において，「背景」および「実施例」の部分の日英対訳文対を対応付けたものであるが，実際に良質な対訳文対が抽出できた部分の割合は約 30%にとどまっている．文献[7]では，「背景」および「実施例」のうちの残りの 70%の部分を言語資源として，既存の対訳辞書を用いた専門用語の訳語推定を行った．本論文では，既存の対訳辞書に加えてフレーズテーブルを用いた結果について報告する．具体的には，NTCIR-7 特許翻訳タスクにおいて配布された対訳特許文対を訓練例として学習したフレーズテーブル，および，既存の対訳辞書に訳語対が登録されていない日英専門用語を対象として，既存の対訳辞書及びフレーズテーブルを用いた要素合成法[6]を適用し，85%以上の高い精度で訳語の推定が可能であることを示す．提案方式を日英対訳特許文書 1,000 文書対に適用したところ，一特許文書対あたりの収集可能な対訳専門用語対の数が，従来方式の平均 3.5 組から平均 4.7 組へと増加した．

2.1.2　日英対訳特許文

本論文では，NTCIR-7 の特許翻訳タスク[1]で配布された約 180 万対の日英文対応データを，フレーズテーブルの訓練用データとして使用した．この文対応データは，1993-2000 年発行の日本公開特許広報全文と米国特許全文を対象として，文献[7]によって日英間で文対応を付けたものである．

2.1.3　要素合成法による訳語推定

2.1.3.1　既存の対訳辞書及びフレーズテーブル

本研究では，既存の対訳辞書として，「英辞郎」[12]に加えて，英辞郎の訳語対から作成した部分対応対訳辞書[5]及びフレーズテーブルを用いる．両者における見出し語数および訳語対数を表 1 に示す．

部分対応対訳辞書生成の手順は以下のとおりである．まず，既存の対訳辞書から，日本語及び英語の用語がそれぞれ 2 つの構成要素(具体的には，日本語の場合は JUMAN[3]による形態素解析によって得られる形態素列，英語の場合は単語列）からなる訳語対を抽出し，これを別の対訳辞書 P_2 とする．次に，P_2 中の訳語対の第一構成要素から前方一致部分対応対訳辞書 B_P を作成し，第二構成要素から後方一致部分対応対訳辞書 B_S を作成する．

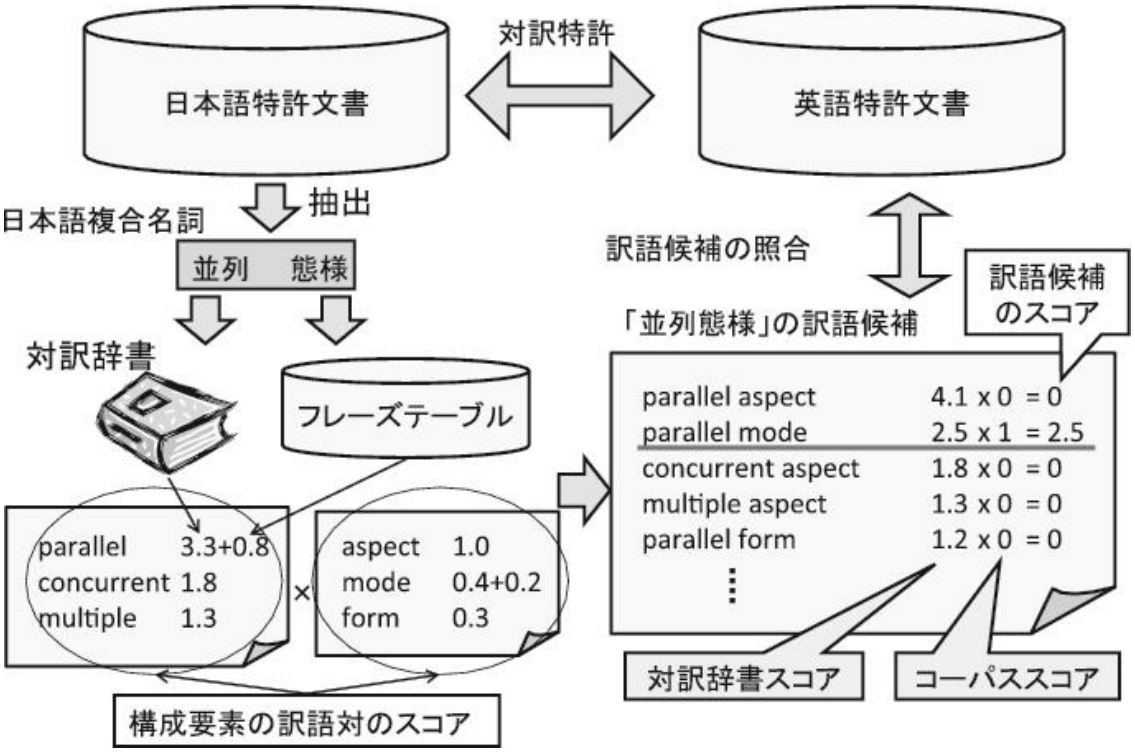


図 1: 日本語の専門用語「並列態様」の要素合成法による訳語推定

[1]http://www.eijiro.jp/

[2]本論文では，英辞郎 Ver.79 及び Ver.131 を用いる．

[3] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

表 1: 英辞郎における見出し語数及び訳語対数

辞書	見出し語数		訳語対数
	英語	日本語	
英辞郎	1,631,099	1,847,945	2,244,117
前方一致部分対応対訳辞書	47,554	41,810	129,420
後方一致部分対応対訳辞書	24,696	23,025	82,087
フレーズテーブル	33,845,218	33,130,728	76,118,632

本論文においては，英辞郎については Ver.131 を使用し，前方一致部分対応対訳辞書及び後方一致部分対応対訳辞書については，Ver.79 及び Ver.131 を統合したものを用いた．

2.1.3.2 訳語候補のスコア

訳語候補のスコアは，対訳辞書スコア $Qdict(y_S, y_T)$ とコーパススコア $Qcorpus(y_T)$ の積で定義される．

$$Q(y_S, y_T) = Qdict(y_S, y_T) \cdot Qcorpus(y_T)$$

ここで y_S は日本語専門用語を，y_T は生成された訳語候補を表し，y_S は構成要素 $s_1, s_2, \cdots, s_n$ に，y_T は構成要素 $t_1, t_2, \cdots, t_n$ に分解できると仮定する．また，対訳辞書スコアはこの構成要素同士のスコアの積によって求まり，コーパススコアは訳語候補が目的言語側のコーパスに生起しているか否かによって求まる．

例として，専門用語“並列態様”の対訳“parallel mode” を獲得する様子を図 1 に示す．本論文では，まず，この日本語専門用語“並列態様” を構成要素 $s1$ の“並列” と $s2$ の“態様”に分解し，これらを既存の対訳辞書及びフレーズテーブルを利用して目的言語に翻訳する．そうすると s_1 からは t_1 として“parallel”,“concurrent”,“multiple”が，s_2 からは t_2 として“aspect”,“mode”,“form” が生成され，さらに各々に訳語の参照元に応じたスコアが付与される．次に，前置詞句の構成を考慮した語順の規則にしたがって，それらの構成要素の訳語を結合し，訳語候補を生成する．このとき，各々の訳語候補の対訳辞書スコアは t_1 と t_2 のスコアの積となる．例えば，“parallel aspect” の対訳辞書スコアは(3.3 + 0.8) × 1.0 = 4.1 である．

最後に，これら訳語候補を対訳辞書スコア順に，目的言語側のコーパスに対して照合を行い，もし照合すればそのコーパススコアは 1，照合しなければ 0 になる．この場合，結果的に，訳語候補のスコアが一番高い“parallel mode” が獲得されることになる．

2.1.3.3　構成要素の訳語対のスコア

構成要素の訳語対〈s, t〉のスコア $q(\langle s,t\rangle)$ は〈s, t〉がどの対訳辞書に出現するかによって場合分けを行った以下の式によって定義される．

$$q(<s,t>)=\begin{cases} & \text{フレーズテーブルの場合} \\ k\cdot P(t\mid s) & (\text{ただし},P(t\mid s)\geq\frac{1}{6}) \\ 10^{(compo(s)-1)} & \text{英辞郎の場合} \\ \log_{10} f_p(<s,t>) & B_p\text{の場合} \\ \log_{10} f_s(<s,t>) & B_s\text{の場合} \end{cases}$$

ここで，$compo(s)$ は s の構成要素数，$f_p(\langle s,t\rangle)$ は，対訳辞書 P_2 中に第一要素として〈s, t〉が出現する回数，$f_p(\langle s,t\rangle)$ は，P_2 中に第二要素として〈s, t〉が出現する回数を表す．また，プレーズテーブルの翻訳確率 $P(t/s)$ をスコアに換算して用いる場合は，換算係数 k の値として $k=1.2$ を用い，$P(t/s)$ の下限値を 1/6 とした場合に訳語推定精度最大となったためこの設定を用いる．

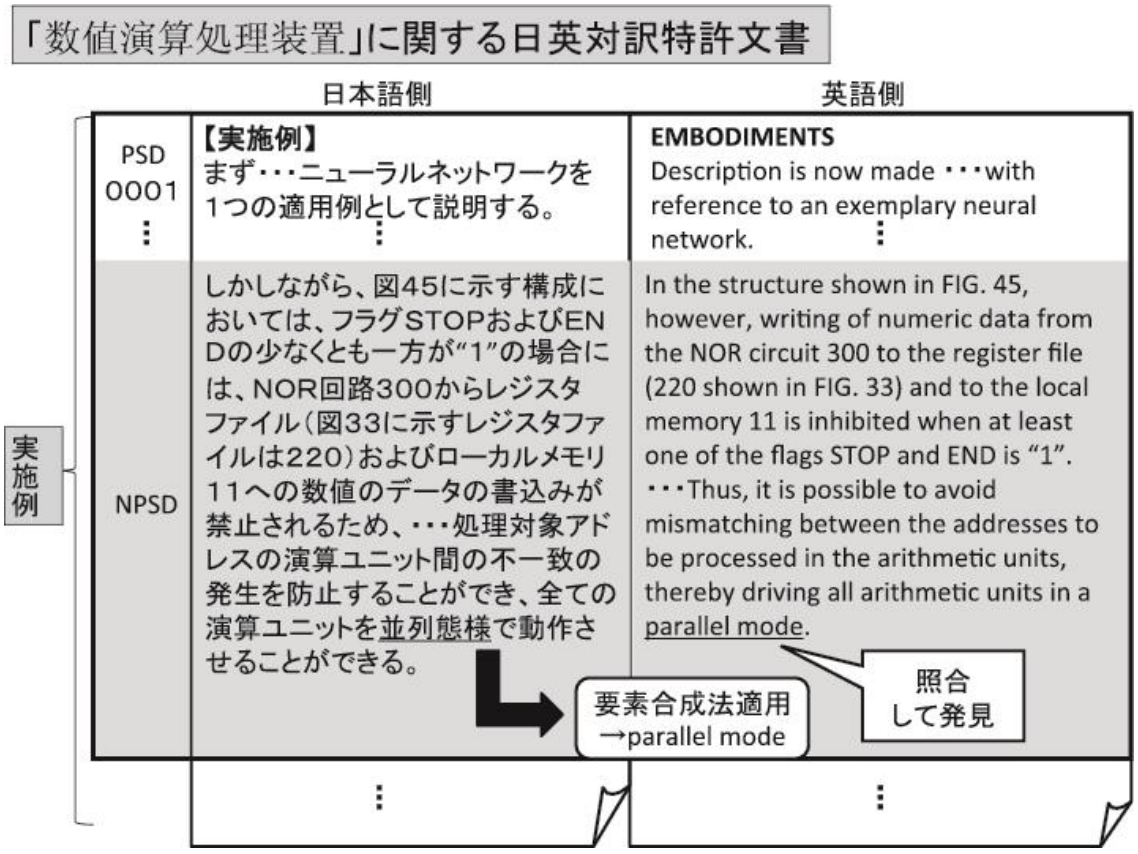


図 2:「実施例」における対訳文対非抽出部分

2.1.4　対訳文非抽出部分における訳語推定

本論文で用いる日英対訳特許文書の日本語側は，「背景」B_J，「実施例」M_J，および，「背景・

実施例以外の部分」N_J から構成されている．そして，これらの部分のうち，「背景」B_J および「実施例」M_J は，対訳文抽出部分 PSD_J，及び，対訳文非抽出部分 $NPSD_J$ に分割される．また，英語側の特許文書の全体 D_E に対しても，同様に，「背景」B_E，「実施例」M_E，および，「背景・実施例以外の部分」N_E から構成され，「背景」B_E および「実施例」M_E は，対訳文抽出部分 PSD_E，及び，対訳文非抽出部分 $NPSD_E$ に分割される．この特許文書の構成の例を図 2 に示す．

$$D_J = \langle B_J, M_J, N_J \rangle$$
$$B_J \cup M_J = \langle PSD_J, NPSD_J \rangle$$
$$D_E = \langle B_E, M_E, N_E \rangle$$
$$D_J \cup M_E = \langle PSD_E, NPSD_E \rangle$$

本論文では，このうちの「背景」B_J 及び「実施例」M_J における対訳文非抽出部分 $NPSD_J$ から日本語専門用語 t_J を抽出した．次に，その日本語専門用語 t_J に対して，英語側の「背景」B_E 及び「実施例」M_E を英語側コーパスとみなして要素合成法を適用し，英語訳語候補の集合 $TranCand(t_J, B_E \cup M_E)$ を作成した[4]

$$TranCand(t_J, B_E \cup M_E)$$
$$= \{t_J \in B_E \cup M_E \mid t_J \text{に対して要素合成法により}$$
$$t_E \text{を生成し} Q(t_J, t_E) > 0\}$$

そして，この $TranCand(t_J, B_E \cup M_E)$ を用いて，以下の関数 CompoTransmax によりスコア最大となる訳語候補を得る．

$$CompoTrans_{\max}(t_J, B_E \cup M_E)$$
$$= \underset{t_E \in TranCand(t_J, B_E \cup M_E)}{\arg\max} Q(t_J, t_E)$$

以上の手順により，日英対訳特許文書の英語側の「背景」B_E 及び「実施例」M_E から英語専門用語 t_E を獲得する．

[4] ここで，比較評価として，英語側の「背景」B_E 及び「実施例」M_E における対訳文非抽出部分 $NPSD_E$ のみを英語側コーパスとみなして要素合成法を適用する評価実験も行ったが，英語側コーパス中において適切な訳語候補を照合できる割合が下がったため，本論文においては，英語側の「背景」B_E 及び「実施例」M_E を英語側コーパスとみなして要素合成法を適用する方式を採用した．

2.1.5 評価

フレーズテーブルを辞書に含めなかった場合，および，含めた場合の 2 通りについて，パテントファミリーである日英対訳特許文書 1,000 文書対を対象として日本語複合名詞を抽出し，その英語訳語を獲得する評価実験を行った．まず，日英対訳特許文書 1,000 組における日本語複合名詞の分類を表 2に示す. 要素合成法の訳語が英語側特許文書中に含まれる日本語複合名詞の数は，フレーズテーブルを辞書に含めなかった場合は 4,060 例，含めた場合は 6,498 例となった．

次に，フレーズテーブルを辞書に含めなかった場合，含めた場合の各々において，要素合成法の訳語が英語側特許文書中に含まれる日本語複合名詞のうち任意の 100 例を抽出し，それぞれ内訳を調査した．

まず，要素合成法の訳語が英語側特許文書中に含まれる日本語複合名詞 100 例を，一般語，評価対象外，専門用語に分類した．この内訳を表 3 に示す．この結果，フレーズテーブルを辞書に含めなかった場合，専門用語は 100 例中 88 例(85%)含まれており，正解であった専門用語は 88 例中 85 例(96.6%) であった．一方，含めた場合，専門用語は 100 例中 84 例(84%)含まれており，正解であった専門用語は 84 例中 72 例(85.7%) であった．ここでの正解とは該当専門用語が日本語特許文書において名詞句として使われており，且つ，その訳語が英語特許文書において名詞句として使われている状態を指す．どちらか一方でも満たしていない場合は不正解とした．また，(i) 接頭辞又は接尾辞が不適切である，(ii) 部分文字列である，(iii) 末尾が識別子である，の場合は評価対象外とした．[5]

表 2: 日英対訳特許文書 1,000 組における日本語複合名詞の分類

分類	件数(割合(%))	
	フレーズテーブル無	フレーズテーブル有
フレーズテーブルの日本語側と完全一致	37,659(61.6)	37,659(61.6)
英辞郎の英訳が英語側特許文書中に含まれる	250(0.4)	240(0.4)
要素合成法の訳語が英語側特許文書中に含まれる	4,060(6.5)	6,498(10.6)
英辞郎または要素合成法により，英訳語候補生成可能であるが英語側特許文書中には含まれない	397(0.7)	551(0.9)
英辞郎または要素合成法により生成不能	18,767(30.8)	16,185(26.5)
合計	61,133(100)	

[5]接頭辞又は接尾辞が不適切とは「上記～，下記～，当該～，該～，各～，～等，～毎」などが接頭辞又は接尾辞に付いている専門用語を指す．部分文字列であるとは，例えば「直角二相変調回路」という全体の文字列の内，部分文字列である「相変調回路」の部分が抽出された専門用語を指す．末尾が識別子とは，例えば「データバッファ装置 DB」のように末尾に「DB」などの識別子の付いている専門用語を指す．

表 3: 要素合成法の訳語候補が英語側特許文書中に出現する 100 例の内訳

分類		件数	
		フレーズテーブル	
		無	有
一般語		0	2
評価対象外		12	14
専門用語	正解	85	72
	不正解	3	12
合計		100	

2.1.6 おわりに

本論文においては，日米パテントファミリーの対応特許文書中において，対訳文が抽出されなかった「背景」および「実施例」のうちの 70%の部分を言語資源として，専門用語の訳語推定を行った結果について報告した．具体的には，NTCIR-7 特許翻訳タスク[1]において配布された対訳特許文対を訓練例として学習したフレーズテーブル，および，既存の対訳辞書に訳語対が登録されていない日英専門用語を対象として，既存の対訳辞書及びフレーズテーブルを用いた要素合成法を適用し，85%以上の高い精度で訳語の推定が可能であることを示した．提案方式を日英対訳特許文書 1,000 文書対に適用したところ，一特許文書対あたりの収集可能な対訳専門用語対の数が，従来方式の平均 3.5 組から平均 4.7 組へと増加した．

参考文献

[1] A. Fujii, M. Utiyama, M. Yamamoto, andt. Utsuro. Overview of the Patent translation task attheNTCIR-7 Workshop. In *Proc. 7th NTCIR WorkshopMeeting*, pp. 389-400, 2008.

[2] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open sourcetoolkit for statisticalmachinetranslation. In *Proc. 45th ACL, CompanionVolume*, pp. 177-180, 2007.

[3] 梁冰, 宇津呂武仁, 山本幹雄. 対訳特許文を用いた同義対訳専門用語の同定と収集. 言語処理学会第 17 回年次大会論文集, pp. 963-966, 2011.

[4] 森下洋平, 梁冰, 宇津呂武仁, 山本幹雄. フレーズテーブルおよび既存対訳辞書を用いた専門用語の訳語推定. 電子情報通信学会論文誌, Vol. J93-D, No. 11, pp. 2525-2537, 2010.

[5] 外池昌嗣, 木田充洋, 高木俊宏, 宇津呂武仁, 佐藤理史. 要素合成法を用いた専門用語の訳語候補生成・検証. 言語処理学会第 11 回年次大会論文集, pp. 17-20, 2005.

[6] 豊田樹生, 高橋佑介, 牧田健作, 宇津呂武仁, 山本幹雄. パテントファミリーを用いた専門用語訳語獲得における対訳文対非抽出部分の利用. 情報処理学会研究報告, Vol. 2012-NL-208, , 2012.

[7] M. Utiyama and H. Isahara. A Japanese-Englishpatent parallel corpus. In *Proc. MT Summit XI*, pp. 475-482, 2007.

[8] V. N. Vapnik. *Statistical Learningtheory*. Wiley-Interscience, 1998.

２．２　複数の２言語辞書とコンパラブルコーパスからの多言語辞書の生成

静岡大学情報学部　梶　博行
綱川隆司
山元陽祐

【要旨】第３言語を介して対訳辞書をマージする方法を一般化し、さまざまな言語対の対訳辞書から多言語辞書を生成する方法を提案する。多言語辞書は、どの２つの要素をとっても対訳であるような対象言語の語の集合をエントリーとする辞書である。対訳辞書のない言語間は他の言語を介してつなぐので、媒介となる語の多義性に起因する誤ったエントリーを除去することが課題である。このためコンパラブルコーパスを用いて出現文脈の類似度を計算し、対訳かどうかを判定する。英日辞書、英中辞書と日中のコンパラブルコーパスを用いた予備評価実験では F 値が最大で 70%であった。コーパスに用例が含まれないエントリーは抽出できないので再現率は低いが適合率は比較的高い。更なる改良が必要であるが、多言語辞書構築のコスト低減が期待される。

2.2.1　はじめに

対訳辞書は、通常、２言語の対訳関係を集めたもので３言語以上を対象にしたものはほとんどない。本研究では、任意個の言語を対象とした多言語辞書を対象言語のうちのいくつかの言語対の対訳辞書とその他の言語対のコンパラブルコーパスから生成する方法を提案する。多言語辞書の各エントリーは対象言語の語の部分集合で、どのエントリーも任意の２つの要素の間に対訳関係が成立するような集合である。１つのエントリーが同一言語の複数の語を含むことはないとする。同一言語の同義語を１つのエントリーにまとめるという考え方もあり得るが、そのようなエントリーは同義語のうちの１つを含む複数のエントリーに分割することができる。

多言語辞書の各エントリーはその要素となっているすべての語に共通の語義を表す。語の多義性は言語によって異なるので、個々の語が多義語であっても、対象言語の数が多くなると語義は限定される (Resnik and Yarrowsky, 2000)。したがって、多言語辞書は個々の言語からみると語義を定義したものとみることができ、語義の曖昧性解消や語義に基づく言語処理に有用な言語資源となる。多言語辞書を利用すれば、複数言語のテキストを入力とする機械翻訳システムを構築することもできる。日本と米国に出願した特許を中国に出願する場合、日本語の明細書とそれに対応する英語の明細書を入力とし中国語の明細書を出力するシステムが考えられ、日英中の３言語辞書を利用することにより訳語選択の精度向上が可能になる。

本研究で提案する方法は第３言語を介した対訳辞書生成方法を拡張したものである。すなわち、２つの２言語以上の対訳語集合が少なくとも１つの語を共有するとき、これらの集合をマージすることによってより多くの言語の語の集合を生成する。ここで、共有されていた語の集合が複数の語義を表す場合、マージして得られる語の集合が対訳語集合であるとは限らない。対訳語集合のみを選択するため、対象言語のコンパラブルコーパスを利用する。すなわち、共有されていなかった語の各々が出現する文脈の類似度が一定の閾値を超える場合のみ、マージして得られる語

の集合が対訳語集合であると判定する。

2.2.2 関連研究

第３言語を介した対訳辞書の生成は Tanaka and Umemura (1994) によって提案された。媒介となる語が多いほど対訳である可能性が高いというヒューリスティクスを用いるものであったが、媒介となる語が１つしかない対訳であることも多く、媒介となる語の多義性の問題を解決するには十分でなかった。Zhang et al. (2007) は、このヒューリスティクスと文字（漢字）の対応に関するヒューリスティクスを組み合わせ、英語を介して日中辞書を生成した。Kaji et al. (2008) はコンパラブルコーパスから語の擬似的な翻訳確率を推定することにより、誤った対訳語候補を除去する方法を提案した。

多言語辞書の生成に関しては、University of Washington の研究グループが２つの方法を提案し、PanDictionary と呼ぶ多言語辞書を構築している。１つは、２言語対訳辞書の集合から翻訳グラフを生成し、その構造から多言語の対訳関係を推定する方法である (Mausam et al., 2009)。もう１つはコンパラブルコーパスを利用する方法である。多言語対訳語集合の候補中のハブ言語（英語）の語を含む文と類似の各スポーク言語の文を検索し、候補中のスポーク言語の語が含まれるとき対訳語集合であると判定する (Sammer and Soderland, 2007)。

「同じ意味の語は似た文脈で使用される」という分布仮説 (Harris, 1954) に基づいて対訳語を抽出する方法はコンパラブルコーパスからの対訳抽出の代表的な方法であり、さまざまな研究が報告されている (例えば Fung and Yee, 1998; Rapp, 1999)。

2.2.3 提案方法

提案方法の前提として、対象言語のうちのいくつかの言語対の対訳辞書が利用可能であり、それらの対訳辞書を連結することによりすべての対象言語がつながるものとする。また、他の言語を介してしかつながらない言語対についてはコンパラブルコーパスが利用可能であるとする。コンパラブルコーパスにはどのレベルのアラインメントも要求しない。同一の分野／ジャンルの単言語コーパスを組にしたものでよいとする。

2.2.1 で述べたように、生成する多言語辞書 D のエントリーはどの２つの要素をとっても対訳であるような語の集合である。したがって、入力となる２言語対訳辞書も対訳の２つの語の集合からなると考える。すべての入力２言語対訳辞書のエントリーの和集合を D の初期値とし、以下の手続きによりエントリーを逐次追加する。

１つ以上の語 $w_1, \ldots, w_K$ を共有する２つのエントリー $\{w_1, \ldots, w_K, \ldots, w_I\}$ と $\{w_1, \ldots, w_K, w'_{K+1}, \ldots, w'_J\}$ が D に含まれ、それらのエントリーの間で共有されない語のすべての組 (w_i, w'_j) $(i=K+1, \ldots, I; j=K+1, \ldots, J)$ が次の(1)または(2)を満たすとき、２つのエントリーの和集合 $\{w_1, \ldots, w_I, w'_{K+1}, \ldots, w'_J\}$ を D のエントリーとして追加する。

(1) w_i と w'_j が対訳であることを入力対訳辞書が示している。すなわち、$\{w_i, w'_j\} \in D$。

(2) w_i と w'_j が対訳であることがコンパラブルコーパスから推定される。すなわち、w_i が出現する文脈 $C(w_i)$ と w'_j が出現する文脈 $C(w'_j)$ の類似度 $Sim(C(w_i), C(w'_j))$ が閾値 θ 以上である。

図１は、英語(EN)、ドイツ語(DE)、日本語(JP)、中国語(CN)を対象言語とし、EN-DE, EN-JP, EN-CN の３つの対訳辞書と DE-JP, JP-CN, DE-CN の３つのコンパラブルコーパスから EN-DE-JP-CN の４言語辞書を生成する場合の例である。*D* の初期値は EN-DE, EN-JP, EN-CN の３つの対訳辞書の和集合である。EN-DE のエントリーと EN-JP のエントリーを組合せ、DE-JP のコンパラブルコーパスを用いることにより EN-DE-JP のエントリーが生成される。同様に、EN-JP のエントリーと EN-CN のエントリーを組合せ、JP-CN のコンパラブルコーパスを用いることにより EN-JP-CN のエントリーが生成される。さらに、EN-DE-JP のエントリーと EN-JP-CN のエントリーを組合せ、DE-CN のコンパラブルコーパスを用いることにより EN-DE-JP-CN のエントリーが生成される。

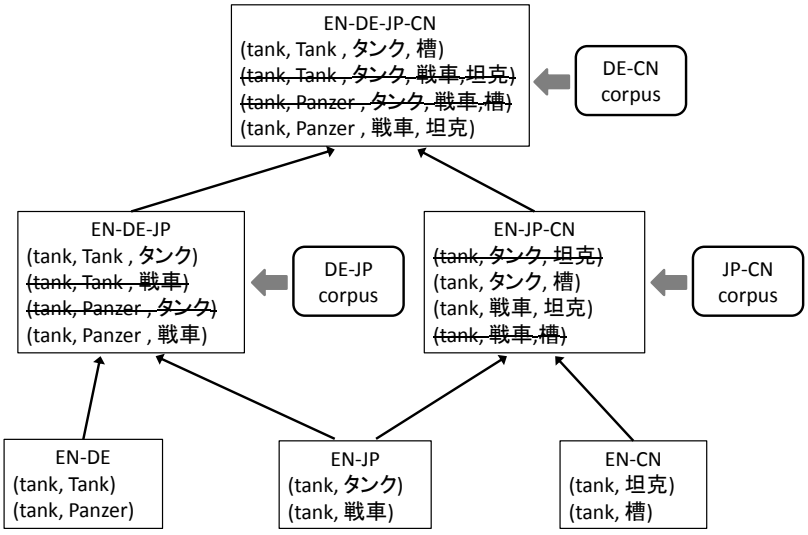


図１　多言語辞書生成方法

１つのエントリーが生成される過程は１とおりとは限らないことに注意されたい。例えば EN-DE-JP-CN のエントリーは EN-DE-JP のエントリーと EN-CN のエントリーを組合せ、DE-CN のコンパラブルコーパスと JP-CN のコンパラブルコーパスを用いることによっても生成される。

2.2.4 コンパラブルコーパスを用いた対訳関係の判定

２つの語が対訳かどうかをコンパラブルコーパスを用いて判定する方法について述べる。

（１）文脈の抽出と表現―重み付き関連語集合

語 *w* の出現する文脈を *w* の関連語の重み付き集合で表現する。関連語の重みとして *w* とのウィンドウ共起に基づく相関値を用いる。*w* と共起する語は非常に多く、また *w* の頻度や性格によって共起する語の数は異なるので、*w* と共起する語のうち *w* との相関値が上位 *p*%の語を *w* の関連語として採用する。

語と語の相関指標としてさまざまなものが考えられる。本稿では対数尤度比(LLR)、χ^2 スコア、相互情報量(MI)、対数オッズ比(LOR)の４つを実験的に比較する。

MI と LOR は低頻度語を過大評価する傾向がある。低頻度語は固有名詞など、後述する文脈類似度計算時に参照する対訳辞書に含まれなかったり、相手言語の対応する文脈に訳語が出現しなかったりするために文脈類似度に寄与しない語が多い。そこで、*w* と共起する語のうち、高頻度語を過大評価する傾向がある LLR の値が上位 *q*%に入り、MI または LOR の値が上位 *r*%に入る語を関連語として採用するバリエーションを追加する（それぞれ LLR&MI、LLR&LOR と略記する。

重みとしてはそれぞれ MI、LOR の値を用いる)。毎日新聞記事コーパスから抽出される“石油”との関連語は、LLR によれば“大手”、“ロシア”、“石炭”、“経済”、…であり、MI によれば“シブネフチ”、“ウルーム”、“ETBE”、“サウジアラムコ”、…である。これに対し、LLR&MI によれば“OPEC”、“採掘”、“ガソリン”、“コンビナート”、…である。この例からも特性の異なる２つの相関指標の組み合わせが有効であると予想される。

（２）文脈類似度の計算―重み付き関連語対応率

異なる言語の文脈の類似度を求めるため種となる対訳辞書を参照する。通常、一方の言語の文脈を翻訳するが、１つの語を対訳辞書が与えるすべての訳語に翻訳せざるを得ないため、翻訳された文脈は多くのノイズを含むという問題がある。多言語辞書の生成ではこの問題が深刻になる。対訳辞書が利用できない言語対の文脈類似度を計算するので、いくつかの言語対の対訳辞書をつないで得られるノイズの多い対訳辞書を参照せざるを得ないからである。例えば、日本語の文脈に“戦車”が含まれているとき、中国語に翻訳すると“坦克”だけでなく“槽”も含まれてしまう。このため、中国語の文脈に“坦克”が含まれている場合でも、文脈類似度は本来の値より小さくなる。そこで、文脈を陽には翻訳しない以下の方法を提案する。

２つの言語の語それぞれについて、「相手言語の語の関連語と対訳関係をもつ関連語の重みの和」と「すべての関連語の重みの和」の比を求め、それらの平均をとる。これを重み付き関連語対応率と呼ぶ。言語１の単語 w の重み付き関連語集合が $C(w)=\{w_i/\alpha_i \mid i=1, \ldots, M\}$ （α_i が w_i の重み、すなわち w と w_i の相関値)、言語２の単語 w'の重み付き関連語集合が $C(w')=\{w'_j/\alpha'_j \mid j=1, \ldots, N\}$（$\alpha'_j$ が w'_j の重み、すなわち w'と w'_j の相関値）であるとき、重み付き関連語対応率 $Sim(C(w),C(w'))$は次式で表される。

$$Sim(C(w), C(w')) = \frac{1}{2}\left(\frac{\sum_{i\in I}\alpha_i}{\sum_i \alpha_i} + \frac{\sum_{j\in J}\alpha'_j}{\sum_J \alpha'_j}\right)$$

ここに、$I=\{i \mid \exists w'_j \in C(w'), (w_i,w'_j)\in D_{12}\}$（$D_{12}$ は言語１から言語２の対訳辞書)、

$J=\{j \mid \exists w_i \in C(w), (w'_j,w_i)\in D_{21}\}$（$D_{21}$ は言語２から言語１の対訳辞書)。

D_{12} や D_{21} がいくつかの言語対の対訳辞書をつないで得られた対訳辞書である場合でも、重み付き関連語対応率が比較的よい類似度指標であることは次の例から理解できよう。対訳の日本語と中国語の語のそれぞれの関連語に“戦車”、“坦克”が含まれる場合、この関連語対が重み付き関連語対応率に寄与することは明らかである。一方の文脈を翻訳する場合と違って重み付き関連語対応率は小さくならない。また、対訳でない日本語と中国語の語のそれぞれの関連語に偶然“戦車”、“槽”が含まれる場合、この関連語対が重み付き関連語対応率を高めてしまう。しかし、対訳でない日本語と中国語の語のそれぞれの関連語集合の間でそのような偶然ばかりが起こることはないので大きな問題にはならない。

2.2.5 予備実験

英日辞書、英中辞書と日中のコンパラブルコーパスを用いて英日中の３言語辞書を生成する実験を行った。実験の目的は重み付き関連語集合の最適案を決定するとともに重み付き関連語対応率の有効性を実証することである。

2.2.5.1 使用データと実験方法

入力の２言語対訳辞書として EDR 日英／英日辞書と LDC Chinese-English Lexicon を使用した。コンパラブルコーパスとして毎日新聞記事（2000 年～2010 年、22.3GB）と LDC Chinese Gigaword の新華社通信記事（2000 年～2010 年、4.24GB）を組にして使用する実験と Wikipedia の日本語版（2012 年 8 月 26 日現在、3.1GB）と中国語版（2012 年 9 月 1 日現在、0.7GB）を組にして使用する実験を行った。

テストデータとして、英語を介して生成された英日中辞書のエントリー候補のうち、日本語と中国語の語のコンパラブルコーパス中の出現頻度が 2,500 以上である候補から 3,000 個をランダムに抽出した。提案方法を適合率、再現率、*F* 値で評価するため、日本語が理解できる３人の中国人学生に全エントリー候補の正誤を判定してもらい、３人の判定結果の多数決により正誤を決定した。3,000 個のうち正（対訳３つ組である）が 1,053 個、誤（対訳３つ組でない）が 1,947 個であった。

重み付き関連語集合には関連語の個数を決定するパラメータ p または (q, r) が含まれる。この値をデータに基づいて決定することが必要であるので、k 重交差検定（k=5）により評価した。すなわち、テストデータの 4/5 を用いて *F* 値を最大化するパラメータ値を決定し、残りの 1/5 のテストデータに対する結果を評価する実験を 5 回行った。なお、ウィンドウサイズも可変パラメータであるが、今回の実験では前後 10 内容語の範囲に固定した。

重み付き関連語対応率の有効性を実証することが１つの目的であるので、ベースラインとして、一方の言語の重み付き関連語集合を翻訳し、重み付き関連語集合をベクトルとみなしてコサイン係数を計算する方法を採用した。したがって、文脈類似度の計算の２つの案（PROPOSED、BASELINE）と重み付き関連語集合の６つの案（LLR、χ^2、MI、LOR、LLR&MI、LLR&LOR）の組み合わせをすべて実行した。重み付き関連語対応率あるいはコサイン係数の降順にテストデータのδ%（δ=10, 20, 30, …）を採用したときの適合率、再現率、*F* 値を算出した。

2.2.5.2 実験結果と結果の検討

新聞記事コーパスを用いた実験の結果を表１に、Wikipedia コーパスを用いた実験の結果を表２にまとめた。スペースの関係で、ベースラインであるコサイン係数と重み付き関連語集合の６案の組合せについては、単一の相関指標（LLR、χ^2、MI、LOR）の中で最もよい結果が得られたものと２つの相関指標の組合せ（LLR&MI、LLR&LOR）の中で最もよい結果が得られたもののみを示す。表の上段が *F* 値で、下段の括弧内は適合率、再現率の順である[1]。各組合せの *F* 値の最大値（イタリック太字）に着目し、８つの組合せをその降順に並べた。*F* 値が最大になる δ の値は組合せによって異なり、また δ の値によって組合せ間の優劣が変わることもあるが、概ね、上に配置されている組合せほど良い結果が得られている。

実験の結果、以下のことが明らかになった。

---

[1] これらの値はテストデータが異なる k（=5）回の実験の平均値である。k 回の実験で決定されるパラメータ値は必ずしも一致しないので、それらの平均を最終のパラメータ値として別のテストデータで評価すべきであるが、テストデータの量の関係で k 回の実験の平均値を評価結果とした。

(1) 文脈類似度の計算方法

異なる言語の文脈の類似度計算方法として、重み付き関連語対応率がベースライン（一方の文脈を翻訳してコサイン係数を計算）より優れていることが確認できた。日本語と中国語の間には対訳辞書（EDR 日中辞書）が利用できるので、これを用いて重み付き関連語対応率とベースラインを比較する実験も行った。表３に示すように、その場合も重み付き関連語対応率がベースラインを上回った。ベースラインとの差は表１や表２のほうが大きく、第３言語を介して作成したノイズの多い対訳辞書しか利用できない場合、重み付き関連語対応率がより効果的であるといえる。

表 1 新聞記事コーパスを用いた実験の結果

文脈類似度	相関指標	δ=10%	20%	30%	40%	50%	60%
PROPOSED	LLR&MI	.410 (.908, .265)	.614 (.833, .486)	***.694*** (.744, .651)	.688 (.640, .746)	.672 (.567, .826)	.636 (.500, .874)
PROPOSED	LLR&LOR	.410 (.908, .265)	.611 (.829, .483)	.687 (.736, .644)	***.693*** (.644, .751)	.672 (.567, .826)	.641 (.504, .881)
PROPOSED	LOR	.403 (.892, .260)	.579 (.788, .458)	.645 (.692, .604)	***.661*** (.615, .716)	.640 (.540, .786)	.611 (.481, .839)
PROPOSED	χ^2	.338 (.767, .217)	.518 (.717, .406)	.587 (.639, .542)	.606 (.571, .646)	***.648*** (.553, .783)	.636 (.506, .858)
BASELINE	LLR&MI	.362 (.783, .235)	.588 (.783, .470)	***.642*** (.678, .610)	.623 (.571, .685)	.600 (.500, .750)	.564 (.439, .790)
PROPOSED	LLR	.380 (.842, .245)	.531 (.721, .420)	.602 (.644, .565)	.621 (.577, .673)	***.627*** (.528, .770)	.608 (.478, .835)
PROPOSED	MI	.368 (.817, .238)	.518 (.704, .410)	.583 (.625, .546)	.612 (.569, .663)	***.618*** (.522, .760)	.615 (.483, .844)
BASELINE	LLR	.346 (.750, .225)	.500 (.667, .400)	***.589*** (.622, .560)	.582 (.533, .640)	.564 (.470, .705)	.568 (.442, .795)

表２ Wikipedia 記事コーパスを用いた実験結果

文脈類似度	相関指標	δ=10%	20%	30%	40%	50%	60%
PROPOSED	LLR&MI	.399 (.883, .258)	.586 (.796, .464)	.666 (.714, .624)	***.700*** (.650, .758)	.678 (.572, .833)	.639 (.503, .878)
PROPOSED	LLR&LOR	.399 (.883, .258)	.571 (.775, .452)	.661 (.708, .620)	***.697*** (.648, .756)	.680 (.573, .835)	.638 (.501, .876)
PROPOSED	MI	.403 (.892, .260)	.577 (.783, .457)	.663 (.711, .622)	***.686*** (.638, .743)	.674 (.568, .827)	.634 (.499, .871)
PROPOSED	LOR	.403 (.892, .260)	.583 (.792, .462)	.663 (.711, .622)	***.684*** (.635, .741)	.658 (.555, .808)	.631 (.496, .866)
BASELINE	LLR&MI	.385 (.833, .250)	.581 (.775, .465)	***.658*** (.694, .625)	.641 (.588, .705)	.632 (.527, .790)	.593 (.461, .830)
PROPOSED	LLR	.365 (.808, .236)	.538 (.729, .426)	.620 (.664, .581)	***.655*** (.608, .709)	.646 (.545, .794)	.620 (.488, .851)
BASELINE	LOR	.362 (.783, .235)	.531 (.708, .425)	.584 (.617, .555)	***.614*** (.563, .675)	.592 (.493, .740)	.582 (.453, .815)
PROPOSED	χ^2	.408 (.883, .265)	.538 (.717, .430)	.563 (.594, .535)	.591 (.542, .650)	***.592*** (.493, .740)	.561 (.436, .785)

表 3 新聞記事コーパスと既存の日中辞書を用いた実験の結果

文脈類似度	相関指標	δ=10%	20%	30%	40%	50%	60%
PROPOSED	LLR&MI	.446 (.967, .290)	.631 (.842, .505)	.700 (.739, .665)	.714 (.654, .785)	***.720*** (.600, .900)	.679 (.528, .950)
BASELINE	LLR&MI	.400 (.867, .275)	.613 (.817, .490)	.674 (.711, .650)	***.700*** (.642, .765)	.672 (.560, .830)	.636 (.494, .890)

(2) 重み付き関連語集合のための相関指標

単一の相関指標を用いるより LLR と MI（あるいは LOR）という異なる特性をもつ相関指標を併用するのが効果的であることが確認できた。単一の相関指標相互の比較では LOR が比較的よかったが、コーパスとの相性もあるようで一概にはいえない。

(3) コーパスのコンパラビリティの影響

新聞記事コーパスの結果と Wikipedia 記事コーパスの結果を比較すると、δ=10～20%では新聞記事コーパスのほうが良いが、δ=40%ではコーパスのサイズが一桁小さい Wikipedia 記事コーパスのほうが良い。文脈類似度を用いる方法ではコーパスのサイズだけでなくコンパラビリティの影響も大きいといえる。

最後に、提案方法の利用に関してコメントする。*F* 値が最大で 0.7 でとても高いとはいえないが、コンパラブルコーパスに用例が含まれないエントリーも多いことを考えるとそれほど悪くもないといえる。再現率はさまざまな分野／ジャンルのコンパラブルコーパスを用いた結果を累積していくことで高めればよいのである。個別のコンパラブルコーパスに対する結果は再現率が 0.25 程度で適合率が 0.9 ほどであるので、その部分を利用することにすれば多言語辞書構築支援ツールとして利用できる。

2.2.6 改良の方向

(1) 語の文脈から語義の文脈へ

文脈類似度を利用する現在の方法には基本的な問題点がある。それは、抽出される文脈はそれぞれの語が表すすべての語義の文脈の総和であり、一つ一つの語義を特徴づける文脈ではないという点である。分布仮説は語義ごとにいえることであり、似た語義を持つ語でも異なる語義で用いられるときの文脈が似ているわけではない。このため、多くの語義をもつ語は訳語との間の文脈類似度がそれほど高くならないことが多い。また、特定の語義で用いられる頻度が高い語はそれ以外の語義での訳語との間の文脈類似度は低くなる。したがって、文脈類似度を利用する方法の性能を高めるには語義ごとに文脈を求めることが必要になる。

１つの方法として、同一の語義を特徴づける関連語どうしの相関が高いことを利用して関連語のクラスタを求めることが考えられる。個々の語がもつ語義の数、そのうちコーパス中に用例が含まれる語義の数とも未知であり、難しい問題であるが、重要な課題である。

(2) 文書レベルのアラインメントが可能なコンパラブルコーパスの利用

2.2.5.2 で述べたようにコンパラビリティの高いコーパスを用いることでよい結果を得ることができると思われる。本稿ではコンパラビリティの低いコンパラブルコーパスにも適用可能な方法を適用しただけであるが、コンパラビリティの高いコーパスに適した方法を開発

することが望まれる。例えば、Wikipedia では記事のアラインメントがとられており、また同一言語の記事の間にもリンクが張られている。これらの情報を積極的に利用する方法によってよりよい結果が得られるであろう。

2.2.7 おわりに

第３言語を介して対訳辞書をマージする方法を一般化し、さまざまな言語対の対訳辞書から多言語辞書を生成する方法を提案した。対訳辞書のない言語間は他の言語を介してつなぐので、媒介となる語の多義性に起因する誤ったエントリーを除去することが課題であった。コンパラブルコーパスから抽出される文脈の類似度によって対訳かどうかを判定する方法として、特性の異なる２つの相関指標を用いて抽出した重み付き関連語集合の間の対応率を計算する方法を開発し、英日辞書と英中辞書から英日中辞書を生成する予備実験によりその有効性を確認した。コーパスに用例が含まれないエントリーは抽出できないので再現率は低いが、適合率は比較的高く、多言語辞書構築のコスト低減が期待される。今後の課題は、語ではなく語義ごとに文脈を抽出するなどの改良と多くの言語への適用評価である。

謝辞：本研究は、一部、文部科学省科学研究費補助金 基盤研究(B)「多義性が解消された多言語辞書の自動構築に関する研究」(課題番号 22300032)の支援を受けた。

参考文献

Fung, Pascale and Lo YuanYee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the ACL and the 17th International Conference on Computational Linguistics*, pp. 414-420.

Harris, Zellig. 1954. Distributional structure. *Word*, 10:146-162.

Kaji, Hiroyuki, Shin' ichi Tamamura, and Dashtseren Erdenebat. 2008. Automatic construction of a Japanese-Chinese dictionary via English. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pp. 699-706.

Mausam, Stephen Soderland, Oren Etzioni, Daniel S. Weld, Michael Skinner, and Jeff Bilmes. 2009. Compiling a massive, multilingual dictionary via probabilistic inference. In *Proceedings of the 47th Annual Meeting of the ACL*, pp. 262-270.

Rapp, Reinhard. 1999. Automatic identification of word translation from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the ACL*, pp. 519-526.

Resnik, Philip and David Yarowsky. 2000. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(3):113-133.

Sammer, Marcus and Stephen Soderland. 2007. Building a sense-distinguished multilingual lexicon from monolingual corpora and bilingual lexicons. In *Proceedings of Machine Translation Summit XI*, pp. 399-406.

Tanaka, Kumiko and Kyoji Umemura. 1994. Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 297-303.

Zhang, Yujie, Qing Ma, and Hitoshi Isahara. 2007. Building Japanese-Chinese translation dictionary based on EDR Japanese-English bilingual dictionary. In *Proceedings of Machine Translation Summit XI*, pp. 551-557.

３．　機械翻訳のための知識獲得－効果的な対訳コーパス構築のための翻訳対象文選択－

京都大学　塩田 嶺明
中澤 敏明
黒橋 禎夫

3.1　はじめに

現在主流となっているコーパスベースの機械翻訳で質の高い翻訳を実現するためには、高品質かつ大規模な対訳コーパスが必要になる。しかし、このような条件を満たし、かつドメインが限られていない対訳コーパスで、広く公開されているものはほとんど存在しない。また、言語によって格差が大きく、話者がそれほど多くない言語では、利用可能なリソースが全くと言っていいほど存在しない場合もある。このため、何らかの方法で対訳コーパスを構築することが強く求められている。人手による対訳コーパス構築の方法として、プロの翻訳家の手によって作成するものの他に、クラウドソーシングを利用するもの[1]などがある。これらの方法に共通する要求は、コスト、翻訳者の労力をできるだけ小さく抑え、効率的にコーパス構築を行いたいというものである。また、コーパス構築にかかるコストを抑えることだけでなく、実際に翻訳システムを構築する際は、計算量をできるだけ抑えるため、目標の翻訳精度をできるだけ少ない数の対訳文で達成するのが望ましい。

人手で対訳コーパスを構築するには、単言語コーパスの文を翻訳していくのが基本となる。単言語コーパスに関しては、様々なドメインにおいて利用可能なものが既に多く存在する他、自力での収集も容易であるので、これを利用すればよい。ここにおける効率的な対訳コーパス構築とは、闇雲に文を翻訳していくのではなく、翻訳された文がどれだけ翻訳システムの品質向上に寄与するかを考慮し、単言語コーパスから翻訳対象となる文を選択して翻訳するというものである。これらの点を踏まえ、本研究では、効率的な文選択手法について考察する。

3.2　能動学習による文選択

文選択においては、能動学習がよく用いられる。能動学習の枠組みを図 1 に示す。まず、元となるデータを翻訳済みデータと未翻訳データに分ける。次に、未翻訳データに含まれる各文に対し、何らかの基準に基づきスコアを付ける。このスコアが上位の文を選択し、翻訳者の手によって翻訳を行う。翻訳された文はその訳文とともに翻訳済みデータに移動される。残された未翻訳データ内の文に対し再びスコアを付ける。このようなサイクルで、対訳コーパスを構築していく。

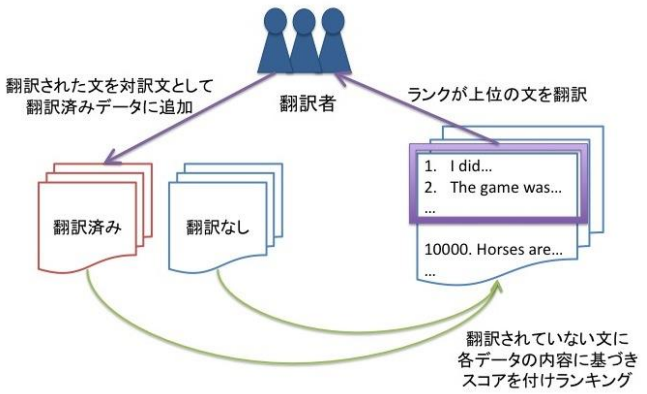


図 1：能動学習の枠組み

3.3　文選択の基準

翻訳対象文を選択する際は、選択された文が翻訳システムの品質向上にどれだけ寄与するか、という点が重要視される。広いドメインをカバーする対訳コーパスを構築するための素朴な発想としては、コーパスが幅広い語彙・フレーズを含むことができるよう、文を選択していくということが考えられる。つまり、未翻訳データ内の文に付与するスコアの指標として、このことを考慮する。Eck ら[2]は、翻訳済みデータ内の文に出現しない単語列 n-gram が文に含まれる数を考慮したスコア指標を提案している。Ambati ら[3]はこれを拡張し、未翻訳データ内での単語列 n-gram の頻度も考慮したスコア指標を提案している。こダムに選択した場合に比べ翻訳精度が向上されることが示されている。Haffari ら[4]は、単語列 n-gram のみに限らず、未翻訳データ内の文を翻訳した精度（ここでは BLEU スコアが用いられている）や、目的言語から原言語に再翻訳した文と元の文の類似度など、様々なスコア指標を能動学習に用いている。しかしやはり、単語列 n-gram など語彙に基づいたスコア指標が、最も良い結果となっている。

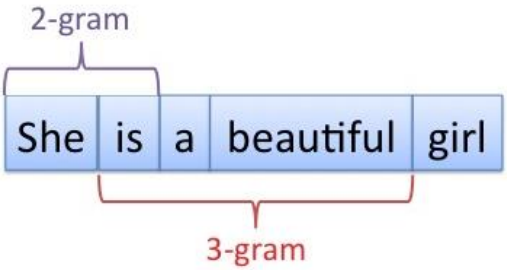


図 2: 単語 n-gram

3.3.1　依存構造木の導入

単語列 n-gram は非常によく用いられる素性であり、能動学においても例外ではない。しかし、
単語列 n-gram は必ずしも意味的にまとまった単位になるとは限らない。例えば図 2 の例文では、”She is”という 2-gram や、”is beautiful”という 3-gram などが得られる。しかし、”is a beautiful”の中だけでは、”beautiful”が指すものが不明であり、”is”との関係が明らかでない。そこで、依存構造木を導入を考える。同じ文を依存構造木で表したものが図 3 である。この構造木上で連続な単語からなる部分木を抽出することを考えると、2 単語からなる部分木”a girl”や、3 単語からなる部分木”She is girl”などが得られる。これらは単語列としては連続ではないが、係り受けの関係にあり、意味的には強いつながりを持っている。このような部分木は、単語 n-gram に代わる文選択のための素性として有用であると考えられる。

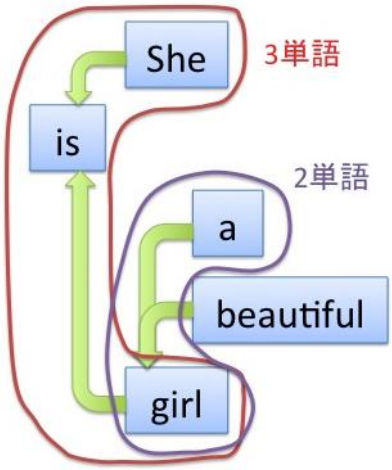


図 3: 依存構造木

表 1: 3-gram の出現頻度

3-gram	出現回数
in order to	16225
be carry out	16218
in this paper	15994
the effect of	12983
this paper describe	10797

表 2: 3 単語の部分木の出現頻度

3 単語部分木	出現回数
be carry out	17840
the effect of	17219
in order to	16002
in this paper	15606
this paper describe	11416

表 1 と 2 は、JST 日英論文抄録コーパスに出現する、単語列 3-gram と 3 単語の部分木の出現頻度を比較したものである。なお、単語は原型に戻してある。例えば、3gram “be carry out”に比べ、同じ単語からなる部分木の出現頻度は多い。これは、例えば”is being carried out”などの場合、”be carry out”からなる部分木が 2 つ含まれるためである。

3.4 能動学習による文選択実験

能動学習による文選択の効果を示すための実験を行った。この実験では既存の対訳コーパスを用いる。仮想的に対訳コーパス構築の過程を再現するため、対訳コーパスのうち一方の言語側はないものとみなし、もう一方の言語側を用いて能動学習を行う。学習過程において選択された文からなる対訳文を用いて、統計的機械翻訳ツールを用いて翻訳システムを構築し翻訳実験を行い、各過程での翻訳精度を調べる。

3.4.1 実験設定

利用したコーパスは JST 日英抄録コーパスである。このコーパスは、科学技術振興機構所有の約 200 万件の日英抄録から、内山・井佐原の方法[5]により、情報通信機構によって作成されたもので、約 100 万対訳文からなる。文選択手法としては、ランダムに選択する方法、能動学習により単語列 n-gram を用いてスコア付けする方法、同じく依存構造木上の部分木を用いてスコア付けする方法の 3 つを用いた。なお、n の最大値は 3 とし、同様に部分木は最大 3 単語までのものとしている。文をスコア付けする式には、Ambati ら[3]が提案している式のうち、以下のものを用いた。

$$Score(S) = \frac{\sum_{Phrase \in S} \alpha}{|Phrase|} \qquad \alpha = \begin{cases} 1 & Phrase \notin Phrase(L) \\ 0 & \end{cases}$$

ここで、S は文を、Phrase は文に含まれる単語列 n-gram または部分木を意味する。L は翻訳済みデータを意味し、Phrase(L)は翻訳済みデータが持つ全ての単語列 n-gram または部分木である。すなわち、このスコアは、文に含まれる単語列 n-gram または部分木のうち、翻訳済みデータに含まれないものの個数を、文全体の単語列 n-gram または部分木の数で割ったものである。

コーパスから 1 万文ランダムに選択したものを初期データとし、能動学習を用いる方法に関しては、一度の学習ステップで 100 文ずつ選択していく。1 文ずつ選択していくことが本来は理想的であるが、計算量が膨大になるため、前述のような設定とした。なお、文選択は対訳コーパスの英語側に着目して行う。このようにして文を選択して作られたコーパスを用いて、10 万文までは 5000 文ごと、10 万文から 20 万文までは 1 万文ごとに翻訳システムを構築した。システムの構築には、統計的機械翻訳ツールである MOSES を用いた。言語モデルとしては SRLIM を利用し、各ステップとも 100 万文全てを使って作成したものを用いた。チューニングは MERT により、500 文を用いて行った。テストデータとして 500 文を用い、翻訳精度評価の指標としては BLEU と RIBES を用いた。

3.4.2 実験結果

実験結果を図 4〜7 に示す。図 4、5 は英→日方向の翻訳の精度、図 6、7 は日→英方向の翻訳の精度をグラフ化したものである。グラフの横軸は翻訳システム構築に用いられた文数を表す。Random はランダム文選択を、String は単語列 n-gram に基づく文選択を、Dependency は依存構造木に基づく文選択をそれぞれ表している。

英→日方向の翻訳に関しては、能動学習を用いた 2 つの文選択手法は、学習過程の初期こそランダム文選択より低い精度に留まっているが、後半に行くに従ってランダム文選択より高い精度を示している。特に依存構造木に基づく文選択手法が高い精度となっている。一方日→英方向の翻訳に関しては、依存構造木に基づく文選択手法とランダム文選択は同程度の翻訳精度となったのに対し、単語列 n-gram に基づく文選択手法はこれらに比べ低い精度に留まっている。文選択は英語側に着目して行ったのにも関わらず、提案手法は日→英方向の翻訳においても効果が見られる。これらを総合すると、依存構造木に基づく文選択手法が最も効率的なコーパス構築への寄与が大きいと言える。

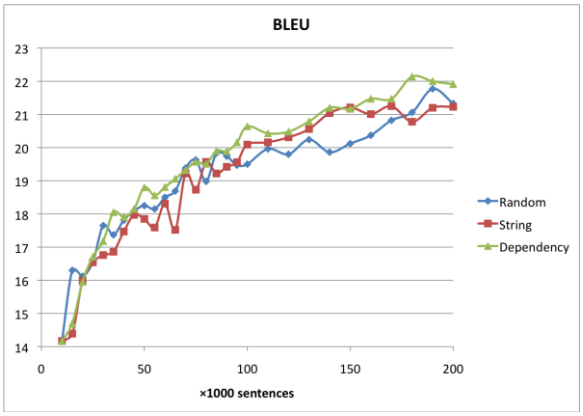


図 4: 英→日方向 BLEU スコア

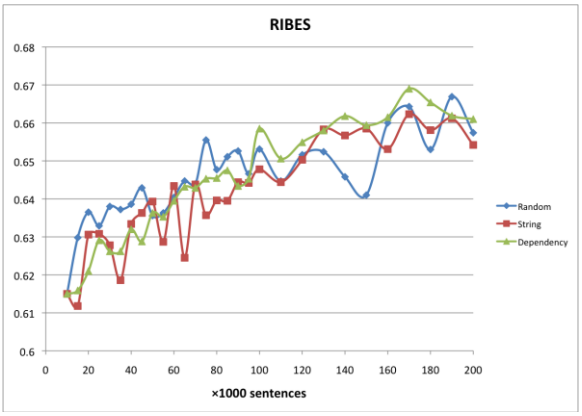


図 5: 英→日方向 RIBES スコア

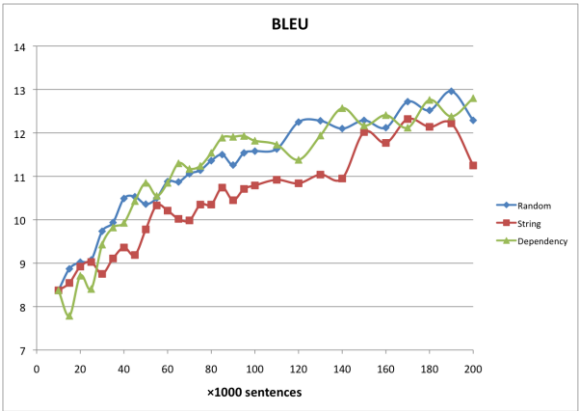


図 6: 日→英方向 BLEU スコア

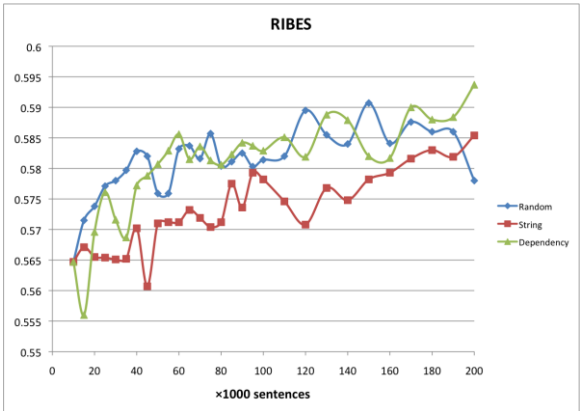


図 7: 日→英方向 RIBES スコア

3.5 おわりに

本研究では、効率的に対訳コーパスを構築するために、単言語コーパスから文を選択する手法について検討した。従来用いられてきた単語列 n-gram に代わり、依存構造木上の部分木を用い

る手法を提案し、既存の対訳コーパスを用いた能動学習による文選択実験を行い、構築された翻訳システムが従来手法に比べ、同じ文数でも高い翻訳精度となることを示した。

今回は能動学習の指針として、翻訳済みデータに出現しない単語列 n-gram または部分木のみに着目し、それを多く含む文を選択していくようにした。今後の方針としてはそれだけでなく、未翻訳データ内の単語列 n-gram や部分木の頻度にも着目して、有用なフレーズを多く含む文を選択していくようにすることが考えられる。

参考文献

[1] Vamshi Ambati, Stephan Vogel, and Jaime Carbonnel. Active Learning and Crowd-sourcing for Machine Translation. Proceedings of the Seventh conference on International Language Resources and Evaluation, pages 2169-2174, 2010.

[2] Matthias Eck, Stephan Vogel, and AlexWaibel. Low Cost Portability for Statistical Machine Translation Based on N-gram Coverage. Proceedings of MTSummit X, pages 227-234, 2005.

[3] Vamshi Ambati, Stephan Vogel, and Jaime Carbonnel. Multi-strategy Approaches to Active Learning for Statistical Machine Translation. Proceedings of MTSummit XIII, pages 122-129, 2011.

[4] Gholamreza Haffari, Maxim Roy, and Anoop Sarker. Active Learning for Statistical Phrase-based Machine Translation. The 2009 Annual Conference of the North American Chapter of the ACL, pages 415-423, 2009.

[5] Masao Utiyama and Hitoshi Isahara. Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp.72–79, 2003.

４．　機械翻訳の分野適応手法に関する調査

東京大学　範　暁蓉
愛媛大学　二宮　崇

4.1　はじめに

大規模な対訳データが利用可能となったことより、統計的機械翻訳の研究が盛んに行われ、統計的機械翻訳はアプリケーションとして供せられる精度を実現しつつある。しかし、特許や新聞、Wikipedia の記事など、特定の分野や特殊な用途における対訳データは大量に存在するものの、多くの分野においては、対訳データがほとんど存在しない、または非常に少ない量の対訳データしか存在しない。大量の対訳データが存在する分野の文書は非常に高い精度で翻訳できるが、少量の対訳データしか存在しない分野では翻訳の質が大きく低下するという問題が知られている。この問題を解決するため、分野適応(domain adaptation)という手法が研究されている。

本稿では、まず、既存のフレーズベース統計的機械翻訳の分野適応に関する手法について述べる。次にこの手法の中からいくつかの手法を選んで、一般分野の学習データを使って、特許分野の分野適応実験をして各手法の性能を評価する。

本稿の構成は以下のようになっている。4.2 節では、分野適応の主要な手法を説明する。4.3 節では、特許分野の分野適応実験について説明する。4.4 節で本稿の主旨をまとめ、今後の課題について述べる。

4.2　分野適応

機械翻訳において、翻訳対象分野のデータ量が十分でない場合や、学習データの分野が翻訳対象の分野と異なるとき、翻訳の精度が低下することが知られている。大量に存する一般分野データと少量の特定分野のデータを併用して、特定分野の翻訳の精度を改善する手法は、分野適応と呼ばれ、さまざまな手法が研究されている。図 1 は、特定分野としての化学分野の対訳データと一般分野としての新聞の対訳データを併用して分野適応を行う様子を示している。

分野適応の研究においては解決するべきことが二つある。一つは特定分野のデータの収集と選択を行うことであり、もうひとつは一般分野のデータと特定分野のデータを併用して学習することである。

4.2.1　特定分野の資源

特定分野は常に以下の四種類の資源がある。

（１）特定分野の対訳用語辞書
（２）特定分野の対訳コーパス
（３）特定分野の用語辞書

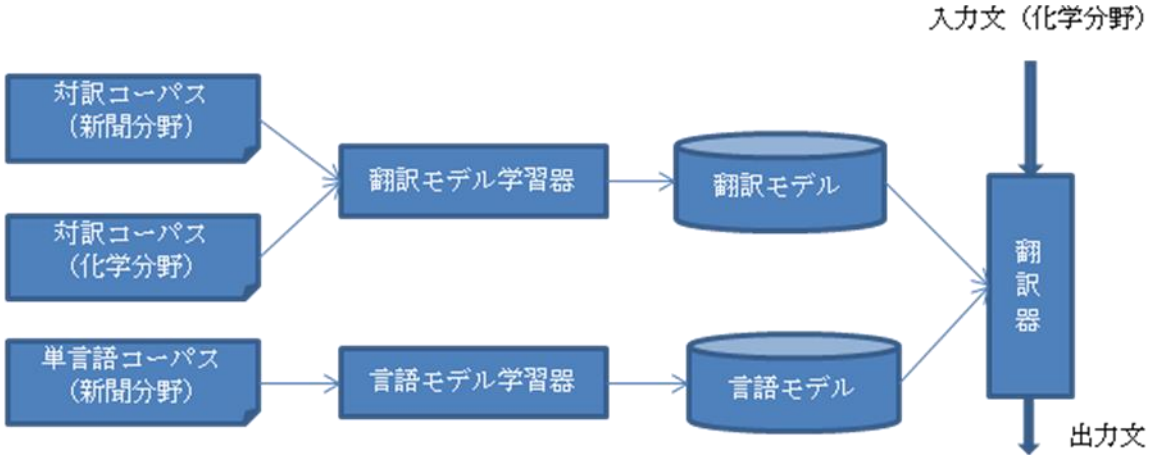


図１ 分野適応時の機械翻訳

（４）特定分野の単一言語コーパス

分野適応を行うとき、この四種類の資源を機械翻訳の異なる段階で利用する。対訳用語辞書は、翻訳モデルか翻訳モデル学習器のリソースとして用いる。対訳コーパスは翻訳モデル学習器のリソースとなる。単言語の用語辞書は使いにくいが、ウェブで訳語をクローリングして、対訳辞書になれば、（１）の資源と同じように用いることができる。単一言語コーパスは二つの用いられ方がある。一つは言語モデル学習器に入れ、言語モデルの性能を改善する。もう一つは、既存の翻訳システムを用いて単一言語コーパスを翻訳することにより訳語を得ることができ、得られた訳語と一緒に翻訳モデル学習器の資源になる。詳しい説明は次節の分野適応手法において述べる。

特定分野の資源収集は一つの大きな研究テーマとなっており、主に三つの手法がある。

（１）手で構築する。

（２）ウェブでクローリングする。

（３）既存の MT システムを使って単一言語コーパスを翻訳して、対訳コーパスを得る。

4.2.2　分野適応手法

図１に示す機械翻訳は四つの部分から成る。（１）学習データ、（２）モデル学習器、（３）学習されたモデルと（４）翻訳器である。データとモデルは機械翻訳のもっとも重要な構成部分である。分野適応の手法はいろいろがあるが、本稿では、まず、データの合成とモデルの合成の二種類の手法を詳しく紹介し、続いて、その他の特別な手法について簡単に説明する。

4.2.2.1　データの結合による分野適応

分野適応は大量に存在する分野外データと分野データを併用することである。一番簡単な方法は分野データと分野外データを結合して一つのデータとする手法である。対訳コーパスの場合、分野対訳コーパスと分野外対訳コーパスを合わせて一つのコーパスとし、それを学習データとして用いて翻訳モデルを学習する。単言語コーパスの場合、分野単言語コーパスと分野外単言語コ

表 1　Philipp Koehn の分野適応の実験の結果

方法	BLEU(%)
Large out-of-domain training data	25.11
Small in-domain training data	25.88
Combined training data	26.69
In-domain language model	27.46
Interpolated language models	27.12
Two language models	27.30
Two translation models	27.64

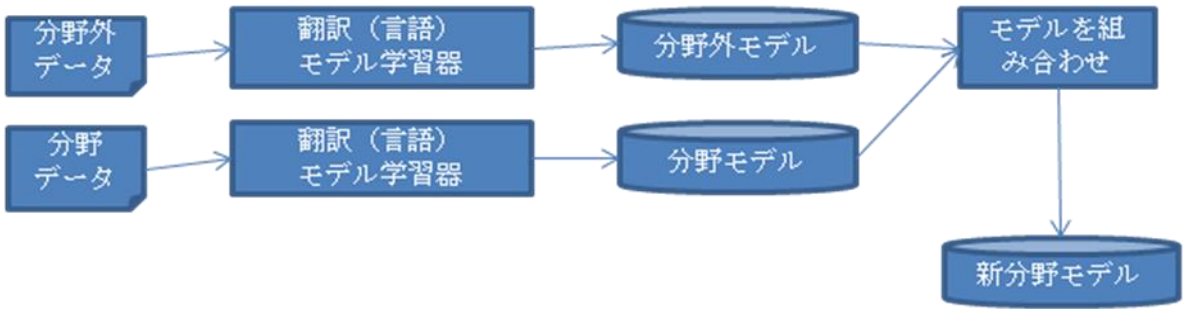


図 2 モデルの合成

ーパスを合わせて一つの単言語コーパスとし、言語モデルを学習する。

この方法の利点は分野適応を容易に行える点である。しかし、この手法による翻訳の改善は小さいことも知られている。Philipp Koehn (2007) らは分野適応の実験をし、表 1 はその結果は示している。Large out-of-domain training data が分野外コーパスだけを使った場合の性能を示しており、Small in-domain training data が分野コーパスだけを使った場合の性能を示している。Combined training data がコーパスを単純に結合する手法の結果を示しているが、この結果を見ると、結合コーパスの結果は単純な分野外コーパスと単純な分野コーパスの翻訳結果より BLEU 値が高くなることがわかるが、これだけでは改善が不十分であることもわかる。分野コーパスの量が少ないことが原因と考えられる。また、In-domain language model は Combined training data による手法において分野単言語コーパスだけを用いて言語モデルを学習した場合の結果となっている。

4.2.2.2　モデルの合成による分野適応

表 1 の結果を見ると、分野データだけを使うと、翻訳結果は単純な分野外データの結果より良いことがわかる。つまり、分野データは分野外データより翻訳の品質に大きい影響があるということが考えられる。しかし、単純なデータの結合による分野適応は、分野データの重要性を表すことができない。この二つのコーパスを同時に利用し、かつ分野データをうまく利用するため、

各コーパスから別々にモデルを学習し、二つのモデルを重みづけして組み合わせる方法が考えられる。図 2 は分野データと分野外データが与えられた時のモデルの組み合わせによる分野適応の様子を表している。以下の節で翻訳モデルの合成方法と言語モデルの合成方法について述べる。

4.2.2.3 翻訳モデルの合成方法

翻訳モデルの合成には、以下の方法がある。

（1） 翻訳モデルの線形補間

線形補間の式を式(1)で示す。

$$p(e \mid f;\lambda) = \sum \lambda_n p_n(e \mid f) \qquad \text{式(1)}$$

$$\sum \lambda_n = 1 \qquad \text{式(2)}$$

ここで、$p_n(e \mid f)$ は学習された各翻訳モデルで、λ_n は各翻訳モデルに対する重みである。

翻訳モデルの線形補間としてはたくさんの手法が研究されている。それぞれの手法において異なるところは重みの設定の仕方である。以下のような重み設定手法が提案されている。

A) Cohn（2007）らは等しい重みを利用した。この手法は簡単だが、分野データの重要性が現れない。
B) Yasuda（2008）らは分野外のモデルの重みを 0.0 から 1 まで変化させ、BLEU 値が一番高かった重み(0.7)を最適な重みとして決定した。
C) Koehn（2010）らはパープレキシティにより、言語モデルの重みを推定した。この重みは翻訳モデルの重みとしても使用された。Rico Sennrich (2012) はパープレキシティによる翻訳モデルの重み設定の方法を提案した。
D) Foster（2010）らは分野外データと分野データの類似度により重みを推定した。

各手法は翻訳精度を改善したが、使用データが異なるため、直接的な比較はできていない。

（2） 対数線形補間

Moses は対数線形モデルも用いた機械翻訳システムである。対数線形モデルは素性として様々な特徴をモデルに組み込みことができる。式(3)は統計的機械翻訳がよく用いる対数線形モデルを示す。

$$p(e \mid f) = \frac{\exp(\sum_n \lambda_n h_n(e, f))}{\sum_{e'} \exp(\sum_n \lambda_n h_n(e', f))} \qquad \text{式(3)}$$

ここで、$h_n(e, f)$ は n 番目の素性を表し、λ_n はその重みである。分野外モデルと分野モデルは二つの素性として結合される。重みは MERT（Minimum error rate training）で推定す

表 2　Foster らの比較実験の結果

Combination	Adapted model		
	LM	TM	LM+TM
Baseline	30.2	30.2	30.2
Loglinear mixture	30.9	31.2	31.4
Uniform linear mixture	31.2	31.1	31.8

る。Philipp Koehn (2007) らの実験はこの補間方法で実験した。表 1 の最後の一行は翻訳モデルの対数補間の結果を示した。結果を見ると、単純なデータの合成より、BLEU 値 (%) が 0.95 上がっている。

Foster (2007)らは線形補間と対数線形補間の比較実験をした。表 2 は彼らの実験結果を示している。この実験における線形補間の重みは等しい重み手法を使った。表 2 によると、線形補間の効果が対数線形補間よりすこし良い。著者である Foster らによると、この実験においては Och の方法では良い極大値を与えるパラメータを学習できていないからではないかと推測している。

4.2.2.4　言語モデルの合成方法

言語モデルの合成方法も線形補間と対数線形補間の二つがある。言語モデルの線形補間の重みはパープレキシティにより重みを設定することが多い。Philipp Koehn (2007)らは分野外のモデルの重みを 0.0 から 1 まで変化して、パープレキシティが一番低い時の重み(0.43 ぐらい)を最適な重みとした。言語モデルの対数線形補間方法は翻訳モデルと同じ補間方法を利用する。

Sanchis-Trilles (2010) らはベイジアン学習方法で式(3)の重みの最適化手法を提案した。この手法は分野コーパスをチューニングデータとし、分野コーパスの量が少ない時、この手法はMERTよりも性能が良い。

4.2.2.5　その他の分野適応手法

Hal Daumé III (2011) らはまず、分野データと分野外データを比較して、未知語を抽出する。次にウェブで未知語の訳語をクローリングする。最後に、得られた分野対訳辞書を現在の翻訳モデルに直接追加する。得られた分野対訳辞書を MT システムに組み入れると分野外のデータを区別ため、四つの新しい素性を追加した。この四つの素性は、（1）クローリング翻訳確率、（2）対訳辞書のペアが分野外データに含むか含まれないかを示す素性、（3）対訳辞書のペアが分野データに含むか含まれないかを示す素性と、（4）2 番目の素性と 3 番目の素性との組み合わせ素性である。

4.3　分野適応実験

特許日英に対し、4.2 節で紹介された手法の一部を使用して、統計的機械翻訳システム Moses を

表 3　実験で用いたコーパスの統計量

コーパス		規模	
		日本語	英語
一般分野	文	487,507	487,507
	単語	11,205,485	10,889,323
	異なり単語数	100,133	92,976
特許分野	文	100,000	100,000
	単語	746,122	656,296
	異なり単語数	16,892	22,850

表 4　分野適応実験の結果

方法	BLEU(%)
Large out-of-domain training data	10.02
Small in-domain training data	11.47
Combined training data	12.50
In-domain language model	10.51
Language models linear mixture	10.95
Language models loglinear mixture	11.05
Two translation models mixture	12.75
Translation models loglinear mixture	12.88

使用して、分野適応実験を行った。使用したコーパスは次のとおりである。

- 一般分野の対訳コーパス：Wikipedia 日英京都関連文書対訳コーパス
- 特許の対訳コーパス：特許データ 2004 年日英タイトル

表 3 はコーパスの基本統計量を示す。

以下の分野適応手法を利用した。

（1） データの結合（Combined training data）

（2） 翻訳モデルの線形補間（Two translation models mixture）

（3） 翻訳モデルの対数線形補間（Translation models loglinear mixture）

（4） 言語モデルの線形補間（Language models linear mixture）

（5） 言語モデルの対数線形補間（Language models loglinear mixture）

表 4 は実験の結果を示す。Philipp Koehn（2007）の実験の結果と異なり、今回の分野コーパスから学習された言語モデルは翻訳精度の改善ができなかった。今回の分野コーパスが特許文書のタイトルからの生成されていること、タイトルが短いこと、主に分野の用語で構成されていることなどが原因のひとつと考えられる。また、他の原因として、特許分野と一般分野の共通部分

が少なかったことが考えられる。モデルの合成は対数線形補間が線形補間よりも良かった。

4.4 まとめ

本稿では、機械翻訳の分野適応手法を調べ、特許分野の分野適応実験を行った。分野適応に有効な手法を考案することは今後の研究課題である。

参考文献

George Foster and Roland Kuhn. 2007. Mixture-Model Adaptation for SMT. In Proceedings of the Second Workshop on Statistical Machine Translation,StatMT '07, pages 128–135, Stroudsburg, PA,USA. Association for Computational Linguistics.

George Foster, Cyril Goutte and Roland Kuhn. 2010. Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 451–459, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hal Daumé III and J. Jagarlamudi. 2011. Domain Adaptation for Machine Translation by Mining Unseen Words, In Proceedings of ACL:Short Papers, pages 407—412, Portland, Oregon, Association for Computational Linguistics.

Keiji Yasuda, Ruiqiang Zhang, Hirofumi Yamamoto and Eiichiro Sumita. 2008. Method of selecting training data to build a compact and efficient translation model. In Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP).

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondˇrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In Proceedings of the Second Workshop on Statistical Machine Translation, StatMT'07, pages 224–227, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn, Barry Haddow, Philip Williams and Hieu Hoang. 2010. More linguistic annotation for statistical machine translation. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR, pages 115–120, Uppsala, Sweden, July. Association for Computational Linguistics.

Sanchis-Trilles, Germán, and Francisco Casacuberta. 2010. Bayesian adaptation for statistical machine translation. Structural, Syntactic, and Statistical Pattern Recognition. 620-629.

Sennrich, Rico, 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of EACL* , pages 539-549.

Trevor Cohn and Mirella Lapata. 2007. Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 728–735, Prague, Czech.

５．　特許文の構造的な特徴－接尾辞に着目した特許文の並列構造解析－

山形大学　横山 晶一

（この原稿は、Japio Year Book 2012 pp.250-253 に掲載したものと同じである。転載を許可された Japio に感謝する）

5.1　はじめに

特許文の構造が、課題や解決手段の部分は複雑であり、200 文字を超える長大な文になるとともに、係り受け構造も複雑であるということは、すでに何度も言及してきた[1-4]。

これまでも、特許文解析に特徴的な、複雑な係り受け構造を解明するため、並列接続詞[5]や、並立助詞[6]について調査し、誤り自動修正システムを構築してきた。

本稿では、並列に重要な役割を果たす名詞を、広く接尾辞としてとらえることによって、特許文の係り受けを修正するシステムについて述べる。ここでの記述は、主として[7]に基づき、その内容に加筆したものである。

5.2　特許文における並列要素

特許文の請求範囲、特に解決手段は、全体が一つの文で書かれることが多く、長大で複雑な文になりやすい。多くは、長い修飾句を有する並列構造になっている。これまで、この修飾句の構造を解明する手がかりとして、並列接続詞、並立助詞について調査し、係り受けの誤りを修正するシステムを構築してきた。本稿では接尾辞に着目するが、それぞれの特徴について簡単に述べる。

(a) 並列接続詞

並列接続詞には、「または」、「もしくは」、「および」、「ならびに」などがある。法律用語では、これらの間に階層関係を設けて、法律の条文の曖昧性をなくす方策を取っている[8]が、我々が特許文を調査した結果では、法律の条文のような階層性は特に見られなかった[5]。

(b) 並立助詞

並列を表す助詞として、「と」、「や」、「か」などがある。特許文では、「A と B との～」といった形でよく使用される。これらの構造をとらえて、係り受けの誤り修正に使用したことがある[6]が、より複雑な構造に対しては、余り効果を発揮できなかった。

(c) 接尾辞

接尾辞とは、一般に、語の末尾に付けて、意味を加えたり、品詞を変化させたり、丁寧さや数など、文法上の変化をもたらしたりする形態素のことを言う[9]。たとえば、「A 部と、B 部と…」における「部」や、「C 層および D 層では…」における「層」が相当する。図１に、特許文の解決手段に書かれた例を示す。この図で、太字で下線を引いた「部」が並列句になっているが、そ

の他の下線の「部」は、この並列構造に寄与していない。この構造は、[]でくくった「と」や「とを」という並立助詞からとらえることができる。

しかしながら、一般的な意味での接尾辞（形態素解析システムでは、「名詞、接尾」といった表示で示される）では、後述するように、特許文の並列構造をすべてとらえることはできない。本稿では、特許文特有の並列構造に係る名詞の概念を拡張して、並列構造に係る名詞を接尾辞としてとらえることによって、並列構造をより明確にすることができることを示す。

[解決手段]電力線通信装置は、電力線に接続された他装置との間で装置情報を送受信する初期接続**部**[と]、前記初期接続部が送受信した装置に基づいて、自装置が通信可能相手情報を作成する通信可能相手情報作成**部**[と]、前記通信可能相手情報作成部が作成した自装置が通信可能相手情報を他装置との間で送受信する通信可能相手情報送受信**部**[と]、前記通信可能相手情報送受信部が送受信した自装置および他装置の通信可能相手情報に基づいて、自装置が通信可能な他装置を階層状にマッピングした登録情報を作成する登録情報作成**部**[と]、前記登録情報作成部によって作成された登録情報を他装置との間で送受信する登録情報送受信**部**[とを]更に設けた。

図１　特許文における並列構造に係る接尾辞の例（公開番号：特開 2010-21954）

5.3　研究方法

5.3.1　資料

解析用の資料として、AAMT/Jspio 特許翻訳研究会特許情報データベース[10]を用いる。このデータベースは、2004 年分の公開特許公報をすべて収録したもので、特許件数約 34 万件、文章数では約 101 万文ある。

5.3.2　接尾辞の抽出と分類

上記資料の文を南瓜[11]に入力して、係り受け解析結果を出力する。図２、南瓜による解析例（一部省略）を示す。図２で、数字の並んだ行は、文節とその係り先を示す。「0 2D」は、「他方部に」が、「畔形成装置と」に係っていることを示す。また、各形態素の右側の記述は、それぞれがどんな品詞となっており、読み方等がどのようになっているかを示している。

図２で、下線を引いた「部」は、既述のように、接尾辞であるが、第２、３文節に現れる「装置」も、並列的な接尾辞と見なすことができる。このような名詞を、本稿では、「特許特有接尾辞」と呼ぶ。

5.3.3　特許特有接尾辞の抽出

本稿では、同じ並列修飾句に並列キーをはさんで並列要素として複数出現する頻度の高い名詞（一般名詞）を「特許特有接尾辞」とする。並列のキー[12, 13]としては、「と」、「や」、「か」、「、」、「・」、「かつ」、「および」、「または」、「ならびに」、「あるいは」、「もしくは」、「だけで（は）なく」を用い、数値とアルファベットを除外して、上記資料の約 101 万文を調査した。

```
他方部に土盛装置と畔形成装置と
*0 2D 1/2 0.644114
他方　　名詞,一般,*,*,*,*,他方,タホウ,タホー
部　　名詞,接尾,一般,*,*,*,部,ブ,ブ
に　　助詞,格助詞,一般,*,*,*,に,ニ,ニ
*1 2D 2/3 0.000000
土　　名詞,一般,*,*,*,*,土,ド,ド
盛　　名詞,固有名詞,地域,一般,*,*,盛,サカリ,サカリ
装置　　名詞,サ変接続,*,*,*,*,装置,ソウチ,ソーチ
と　　助詞,並立助詞,*,*,*,*,と,ト,ト
*2 -1D 2/3 0.000000
畦　　名詞,一般,*,*,*,*,畦,アゼ,アゼ
形成　　名詞,サ変接続,*,*,*,*,形成,ケイセイ,ケイセイ
装置　　名詞,サ変接続,*,*,*,*,装置,ソウチ,ソーチ
と　　助詞,並立助詞,*,*,*,*,と,ト,ト
EOS
```

図２　南瓜の解析例

並列句に現れる高頻度名詞（339716 特許、1013582 文）		
手段	名詞,一般	23523 回
装置	名詞,サ変接続	19906 回
工程	名詞,一般	12305 回
方法	名詞,一般	10683 回
情報	名詞,一般	7229 回
基	名詞,一般	6579 回
データ	名詞,一般	4671 回
部材	名詞,一般	4534 回
ステップ	名詞,一般	3967 回
位置	名詞,一般	3674 回

図３　特許特有接尾辞の一部

図３に、抽出した特許特有接尾辞の一部を示す。本研究では、上位 100 語、出現頻度 413 回以上の語を用いる。

5.3.4　分野別の出現頻度

特許には、国際的に用いられている文献の技術内容によるカテゴリがあり、記号によって分類されている。「セクション」、「サブセクション」、「クラス」、「サブクラス」、「メイングループ」。「サブグループ」の順に分類が細かくなっている。

セクションは、A（生活必需品）、B（処理操作、運輸）、C（化学、冶金）、D（繊維、紙）、E（固定構造物）、F（機械工学、照明、加熱、武器、爆破）、G（物理学）、H（電気）の８分野に分かれている。

図４に、セクション C の特許データについて調査した結果を示す。すべての特許文を対象にした場合に高頻度にならなかった「原子」、「酸」などの名詞が上位に来ていることが分かる。

化学、冶金分野の高頻度名詞（27969 特許、76517 文）		
基	名詞,一般	3729 回
以下	名詞,非自立,副詞可能	1941 回
工程	名詞,一般	1844 回
樹脂	名詞,一般	1654 回
方法	名詞,一般	1613 回
原子	名詞,一般	633 回
酸	名詞,一般	580 回

図４　C(化学、冶金)分野における特許特有接尾辞

5.4　係り受け修正システム

特許特有接尾辞を用いたシステムを作成すると、従来はできなかった、並列構造の中に入れ子構造的にまた並列構造を含む文の係り受け誤りを修正することができる。

図５に、このようなシステムの出力例を示す。図の左側の番号は、文節番号を示し、その右側の 2D, 3D という番号は、係り先の文節を示す。係り受けの修正を行った場合には、(8 17D)のようにカッコに入れて表示し、係り元の接尾辞を[ ]で、係り先の接尾辞を< >で囲んだ。

この図の文は、文節 8 と文節 17 の並列構造の中に、文節 9 と文節 17 の並列構造を入れ子的に含んでいる。このような文でも、特許特有接尾辞（ここでは「装置」）に着目することにより、係り受けを修正することができる。

5.5　評価と考察

2004 年分のデータから、接尾辞を含む並列構造文 500 文を無作為に選び、システムにかけた結果、どの程度誤りを修正することができたかを評価した。その結果を表１に示す。

この表で、「正→正」は、もともと並列構造を正しく解析できていたものをこのシステムにかけてもやはり正しく解析できたことを示し、「誤→正」は、解析誤りを修正できたことを示す。

表１からわかるように、誤り 166 件（97＋69 件）のうち、97 件(58.4%)を修正できた。これは

過去の結果[4, 5]の 55.7%よりもわずかによい結果となっている。

0 28D		また、
1 2D		その
2 3D		ための
3 28D		制御冷却装置は、
4 5D		圧延直後の
5 6D		鋼板の
6 7D		麺温度分布を
7 8D		測定する
8 15D	(8 17D)	温度測定[装置]と、
9 11D		冷却水ヘッダーと
10 11D		これに
11 12D		接続された
12 13D		ラミナー状の
13 14D		冷却水を
14 15D		供給する
15 16D		ノズルとを
16 17D		含む
17 23D	(17 27D)	冷却<[装置]>と、
18 19D		所定の
19 20D		計算プログラムに
20 23D		したがって
21 22D		鋼板の
22 23D		麺温度分布を
23 25D		均一化するように
24 25D		冷却水量を
25 26D		制御する
26 27D		冷却水量の
27 28D		制御<装置>とを
28 -1D		備える。

図５　システムによる係り受け誤りの修正成功例

表１　システムによる係り受け誤り修正結果

	正→正	正→誤	誤→正	誤→誤	合計
件数	318	16	97	69	500
%	63.6	3.2	19.4	13.8	100

5.6　まとめと今後の課題

本研究では、接尾辞を考慮することによって、並列構造の解析誤りを修正する試みについて述べた。特許特有接尾辞の数を限定したことが、誤り修正の程度が期待ほど上がらなかった原因と考えられる。今後は、まず、語数の増加をはかる。次に、シソーラスを組み合わせたり、接尾辞の直後に来る数値や記号に着目したりすることによって、さらに精度の向上を目指す予定である。

謝辞

特許データベースをご提供いただいた Japio に感謝致します。

参考文献

[1] 横山晶一、高野雄一：語のグループ化を用いた特許文動詞の自動訳し分けに関する調査、Japio Yearbook (2011) pp.234-237

[2] 横山晶一、高野雄一：特許文の英語への訳し分けと述語の関係、Japio Yearbook (2010) pp.274-279

[3] 横山晶一：特許文の英語への訳し分けと格フレームとの関係、Japio Yearbook (2009) pp.262-265

[4] 横山晶一：動的シソーラスを用いた特許文の解析システム、科学技術研究費成果報告書(2007～2009)

[5] 横山晶一：特許文における接続詞と係り受けの構造、Japio Yearbook (2008) pp.68-73

[6] 横山晶一：特許文解析誤り自動修正システムと正確な翻訳のための特許文の分割、Japio Yearbook (2007) pp.228-233

[7] 坂本和磨：接尾辞に着目した特許文の並列構造解析、山形大学工学部情報科学科卒業論文(2012)

[8] 田島信威：最新法令用語の基礎知識（三訂版)、ぎょうせい(2006)

[9] 新村出編：広辞苑（第六版)、岩波書店(2008)

[10] AAMT/Japio 特許翻訳研究会特許データベース(2004)

[11] 奈良先端科学技術大学院大学：係り受け解析器「南瓜」

[12] 山村広臣、菅沼明、牛島和夫：日本語文における並列構造の簡便な推定法および推敲支援への適用、情報処理学会第 52 回全国大会(1997)

[13] 岩本秀明、長野馨、永井秀利、中村貞吾、野村浩郷：法律文における並列構造の特徴とそれに基づく制限言語モデルについて、情報処理学会自然言語処理研究会(1993)

6．1　機械翻訳の評価について

山梨英和大学　江原 暉将

6.1.1　はじめに

技術の進歩には正当な評価が必要であり、機械翻訳においてもそうである。しかしながら翻訳という極めて人間的な作業を正当に評価すること自体困難なことである。機械翻訳の評価は、翻訳結果を人間が評価する「人手評価」から始まり、近年 BLEU をさきがけとする「自動評価」が発展してきた。しかし人手評価にも自動評価にも、まだまだ課題が多い。当研究会ではこれまでも評価に関する研究や調査を行ってきたが、今年度は特に外部の識者を招き、また一般からの参加も得て、「特許文書の機械翻訳結果評価方法検討会」を開催した。さらに当研究会委員および外部専門家にも加わってもらい「拡大評価部会」を立ち上げた。本報告では「検討会」での議論と「拡大評価部会」の活動概要を記す。

6.1.2　特許文書の機械翻訳結果評価方法検討会

2012 年 9 月 7 日に東京大学工学部 11 号館講堂において表題の検討会を開催した。参加者は 96 名で活発な議論が展開された。

本検討会での議論の焦点は以下の通りである。

(1) 外国の特許文書からその技術内容を調査する場面での機械翻訳の評価
(2) 自動評価の現状、そして、利用者の評価にできるだけ近い自動評価方法の探索
(3) 人手評価と自動評価の比較
(4) 特許翻訳用テストセットの要件
(5) 今後の翻訳評価の方向性

プログラムを以下に示す。講演資料は当研究会のホームページに掲載してある。

1. 自動評価方法の研究者の立場から
   北海学園大学　工学部生命工学科　越前谷博氏
   講演題目：自動評価がもたらした歓喜と失望、そして、希望
   岡山県立大学　情報工学部情報システム工学科　磯崎秀樹氏
   講演題目：最近の自動評価法の研究動向と RIBES
2. 機械翻訳メーカの立場から
   (株)東芝　研究開発センター 知識メディアラボラトリー　鈴木博和氏
   講演題目：「空気の読める機械翻訳」の評価方法
3. 人手評価の立場から
   情報通信研究機構　後藤功雄氏　(NTCIR-10　PatentMT タスクオーガナイザ)
   講演題目：NTCIR-9，NTCIR-10 特許機械翻訳タスクでの人手評価
4. 企業の技術調査担当者の立場から
   トヨタテクニカルディベロップメント(株)　森田陽介氏　(知的財産情報検索委員会副委員長)
   講演題目：特許調査に求められる機械翻訳の精度～中国特許調査の事例より～
5. 評価用テストセット作成の立場から

(株)富士通研究所　メディア処理システム研究所　長瀬友樹氏　(AAMT機械翻訳課題調査委員会委員長)

講演題目：テストセットを用いた日中翻訳エンジン評価

6. 総合討論　(今後の翻訳評価方法の展望)

議論の概要を箇条書き的にまとめて示す。

・特許文書の機械翻訳結果の評価といっても「評価する人は誰か」「評価の目的は何か」によって評価方法が変ってくる。
・評価する人としては、MT の研究者、MT の開発者、MT の利用者がある。
・特許分野で MT の利用者が評価する場合、評価の目的としては、技術調査、外国への出願、特許庁での審査などがある。
・評価のダイナミックレンジと解像度を考えないと評価方法も決められない。
・技術調査目的では訳語の正確さが第一に重要である。
・一般的な評価としては訳語より文法的正しさがまず必要、MT のレベルに合わせて評価方法も変化するべき。
・人手評価としては Adequacy(訳語の正確さ)、Fluency(文法的正確さ)に対応するのではないか?
・自動評価としては RIBES(文法的正確さ、語順の正確さ)に対して BLEU(訳語の正確さ)が対応するか?　IMPACT は双方を重視しているか?
・一般的評価ではなく、利用場面を限った評価も有用。
・開発者がシステムのエラー分析を行うためにはテストセットと設問ベースの評価が有用。
・特許審査場面での評価では NTCIR-10 で採用する「特許審査評価」が有用。
・「空気の読める機械翻訳」といったかなり高度な場面での評価では人間の ESOL 用テスト基準(CEFR 準拠)が有用。
・言語知識や言語固有の現象を評価に組み入れる必要がある。
・日本語固有の現象を組み入れてはどうか。
・言語固有の知識をパラメータとして取り入れられると良い。
・自動評価に望まれる条件としては以下がある。
  ・低コスト(reference は 1 個または 0 個が望ましい、計算が軽い)
  ・SMT の Tuning に利用可能
  ・スコアの意味が明確
  ・人手評価との相関が高い
  ・言語依存のリソースが不要
・特許用の機械翻訳テストセットが作れると良い。
・特許用のテストセットと設問を作成できる可能性はある。
・作成コスト削減のために特許ファミリーを利用するべき。
・テストセットを作るのならば、分野(電気、機械、化学など)と特許文書構造(請求項、実施例など)別に作るべき。

6.1.3　拡大評価部会の設置

「特許文書の機械翻訳結果評価方法検討会」での議論の結果、評価に関してまだまだ課題が多いことが明らかとなった。そこで、当研究会委員および外部専門家にも加わってもらい「拡大評価部会」を立ち上げた。拡大評価部会は AAMT/Japio 特許翻訳研究会の下部組織であり、部会員(平成 24 年度)は以下の通りである。部会長は江原が担当している。

当研究会委員：

辻井潤一、横山晶一、江原暉将、宇津呂武仁、越前谷博、後藤功雄、潮田明、須藤克仁

外部専門家：

磯崎秀樹　岡山県立大学教授

鈴木博和　(株)東芝　研究開発センター　知識メディアラボラトリー

長瀬友樹　(株)富士通研究所　メディア処理システム研究所

Japio 関係者：

大塩只明、王向莉

本部会での研究の焦点は以下の通りである。

1. 「技術調査目的」のために特許文書を機械翻訳する場合の評価
2. 人手評価、自動評価、半自動評価
3. 評価用テストセット
4. 対象とする言語の範囲　日本語、英語、中国語
5. 評価手法の理想形、理想を実現するための課題、課題克服への道程

本年度は部会を 1 回開催し、前記「検討会」の議論を深めた。その結果、今後、自動評価、人手評価、テストセットについて研究を進めることとなった。また、部会員より本章に 5 編の寄稿をいただいた。さまざまな視点から機械翻訳の評価を多角的に捉えた章となっているものと思う。

6.1.4　まとめ

特許文書に関する機械翻訳の評価について、「特許文書の機械翻訳結果評価方法検討会」と「拡大評価部会」の概要を報告した。今後、「拡大評価部会」を中心に機械翻訳のより良い評価手法の実現を目指して研究を進めたい。

６．２　自動評価の現状と今後の方向性について

北海学園大学　越前谷　博
岡山県立大学　磯崎　秀樹
NTT コミュニケーション科学基礎研究所　須藤　克仁

6.2.1　はじめに

機械翻訳の評価において自動評価に大きな期待が寄せられている。しかし、現時点ではまだ十分なものが提案されるまでには至っていない。そこで、これまでに提案された代表的な自動評価尺度を整理し、新たな自動評価尺度の方向性を考える際の足掛かりとする。

自動評価に関する取り組みは 2002 年に提唱された BLEU[1]を出発点として急速に加速した。これは統計翻訳の研究が活発に行われるようになったことを背景として、機械翻訳の評価をどのように実施するのかがこれまで以上に切実な問題となったためである。その際、人手による評価は高い精度を期待できるが、時間とコストがかかる。また、評価者ごとの揺れが生じるため、客観的な評価という点で問題がある。このような問題点を克服するために自動評価の研究が盛んに行われるようになり、これまでに様々な自動評価尺度が提案されてきた。そこで、自動評価がどのような変貌を遂げてきたかを改めて振り返り、より良い自動評価について考える。

6.2.2　理想的な自動評価

Philip Koehn[2]は 3 つの観点より自動評価に求められるものについて言及している。1 つ目は low-cost metrics である。これは自動評価に求められる必須要件である。人手評価と同じようにコストと時間を要するものであっては、自動評価を用いる意味が失われる。2 つ目は tunable metrics である。統計翻訳においては自動評価がチューニングとしての役割を担っていることから、不可欠な要件として取り上げられている。3 つ目は meaningful metrics である。基本的に自動評価尺度は MT 訳文に対して評価結果として数値を与えるが、その数値がどのような意味を持つのかが明確でなければならない。

これらの 3 つの要件に加え、更に以下のような要件が自動評価に必要と考えられる。

・人手評価との相関が高い
・特定の言語に依存しない
・処理時間が短い
・参照訳が少ない

自動評価に求められるものは使用者及び開発者によって異なる。しかし、いずれの立場であっても人手評価との相関が高いこと、すなわち、高い評価精度を有していることが求められる。これを満たしていなければ他の要件を仮に満たしていたとしても、その存在意義は損なわれることになる。また、その他には、統計翻訳とルールベースなどの異なるアーキテクチャーを用いた機械翻訳システムに対しても、アーキテクチャーに依存することなく正しく評価できることが要求される。

6.2.3　スタンダードな自動評価尺度

現在、最も広く利用されている自動評価尺度は BLEU である。BLEU は下記の式よりスコアを得る。スコア値は 0.0 から 1.0 の範囲で出力され、1.0 に近いほど評価対象の MT 訳文は良い訳文と評価される。BLEU は n-gram 適合率に基づく手法である。p_n が MT 訳文における適合率を表している。n の値としては 1 から 4 までを用いるのが最も良いとされている。また、MT 訳文の単語数が小さくなると、参照訳と大きく異なっていても評価が高くなる。そのため、BP をペナルティとして重み付けに用いている。そして、n を 1 から順に変化させた際の適合率の幾何平均を求めることで BLEU スコアを算出する。BLEU は処理が非常にシンプルであり、処理時間も高速である。しかし、BLEU は n-gram の幾何平均を用いているため、ある n-gram の値が 0 になると最終的なスコア値も 0 になる。このような現象は単語数の少ない文単位では高い確率で発生することから、文単位の評価には適さないことが知られている。

$$p_n = \frac{\sum_{c\in\{Candidates\}}\sum_{n-gram\in C} Count_{clip}(n-gram)}{\sum_{C'\in\{Candidates\}}\sum_{n-gram'\in} Count(n-gram')}$$

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \le r \end{cases}$$

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

また、相互情報量により重み付けされた n-gram 適合率に基づく手法として NIST[3]も広く利用されている自動評価尺度であり、以下の式より得られる。NIST ではスコア値は 0.0 以上であり、値が大きいほど高い評価であることを表す。また、高速処理が可能であるが、BLEU と同様に文単位の評価には適さない。

式 *Info* は n-gram 適合率を相互情報量で表している。例えば、単語”the”が 10 回出現した場合、”the”に続く単語として”cat”が 9 回、”dog”が 1 回出現していたとすると、式 *Info* の値は”the cat”よりも”the dog”の方が高くなるため情報量が多くなる。このように情報量を用いることで意味にも考慮したスコアとなる。式 *Score* では、得られた相互情報量をそのまま重み付けとして用いている。また、*exp* は BLEU と同様にペナルティを表している。MT 訳文の単語数が参照訳の単語数よりも大きければ 1、その逆であれば 1 以下が選択される。

$$Info(w_1 \dots w_n) = \log_2\left(\frac{the\ \#\ of\ occurances\ of\ w_1 \dots w_{n-1}}{the\ \#\ of\ occurrences\ of\ w_1 \dots w_n}\right)$$

$$Score = \sum_{n=1}^{N}\left\{\frac{\sum_{all\ w_1\dots w_n that\ co-occur} Info(w_1 \dots w_n)}{\sum_{all\ w_1\dots w_n in\ sys\ output}(1)}\right\} \cdot exp\left\{\beta \log^2\left[min\left(\frac{L_{sys}}{\bar{L}_{ref}}, 1\right)\right]\right\}$$

BLEU や NIST が文単位の評価に適さない自動評価尺度であるのに対して、文単位の評価を重視した自動評価尺度として METEOR(Metric for Evaluation of Translation with Explicit Ordering)[4]が広く知られている。METOR では、文単位での評価精度を向上させるために、適合率だけでなく再現率も用いている。METEOR は以下の式より得られる。METEOR のスコア値は 0.0 から 1.0 であり、値が大きいほど高い評価となる。また、METEOR では、オプションとして、文字列のマッチングだけでなく単語の語形変化や WordNet による類義語を用いた評価を行うことができる。したがって、文単位の評価に有効である。しかし、処理時間がかかる。

$$F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R}$$

$$Pen = \gamma \cdot \left(\frac{ch}{m}\right)^{\beta}$$

$$score = (1 - Pen) \cdot F_{mean}$$

式 F_{mean} の P と R はそれぞれ適合率と再現率を表している。適合率とは MT 訳文の単語数に対する一致単語の割合であり、再現率とは参照訳の単語数に対する一致単語の割合である。この P と R の F 値が F_{mean} である。式 Pen はペナルティである。m は一致単語数を、ch はチャンク数である。例えば、MT 訳文と参照訳が完全に一致している場合、ch は 1 となりペナルティは小さくなる。一方、個々の単語が完全一致ではあるが、語順が逆順になっている場合、$ch=m$ となりペナルティは大きくなる。そして、式 $score$ ではこの Pen と F_{mean} を用いてスコア値を求めている。Pen が小さい場合にはスコア値は大きくなり、Pen が大きい場合にはスコア値は小さくなる。変数 α、β、γ はパラメータである。

更に最もシンプルな自動評価尺度として WER（Word Error Rate）[5]がある。WER はレーベンシュタイン距離、すなわち編集距離に基づく尺度である。MT 訳を参照訳に一致させるために、substitutions（置換）、insertions（挿入）、deletions（削除）の 3 つの操作を何回実行すればよいかを求めることで得られる。以下にその式を示す。WER のスコア値は 0.0 以上で値が小さいほど評価が高い。処理速度は比較的高速であり、文単位の評価にも有効である。また、特徴として、語順がスコアに強く反映されることが挙げられる。

$$WER = \frac{substitutions + insertions + deletions}{reference - length}$$

6.2.4　準スタンダードな自動評価尺度

6.2.3 で述べた自動評価尺度ほど広く利用されてはいないが、自動評価の研究においてよく取り上げられる自動評価尺度について述べる。これらは、比較実験の際にベースラインとして用いられていることが多い。

ROUGE シリーズ[6]では 3 つの自動評価尺度が提案されている。1 つは Longest Common Subsequence に基づく ROUGE-L である。ROUGE-L は参照訳の単語数に対する LCS の長さの割合を再現率、MT 訳文の単語数に対する LCS の長さの割合を適合率として、再現率と適合率の F 値を求め、それをスコア値とするものである。2 つ目は ROUGE-L の拡張版である、Weighted Longest Common Subsequence に基づく ROUGE-W である。ROUGE-W は、共通部分の長さに応じた重み付けを行っており、ROUGE-L と同様に再現率と適合率の F 値をスコア値としたものである。そして、3 つめは、n-gram に対して、語順の連続性の制約を排除した一致率（Skip-bigram）に基づく ROUGE-S である。例えば n=2 の bigram であれば 2 つの単語が連続して一致していなければならないが、skip-gram では、連続している必要がない。その場合、一致率の分母は 2 つの単語の組み合わせ数を用いる。これらの ROUGE ではスコア値は 0.0 から 1.0 であり、値が大きいほど高い評価となる。ROUGE-L と ROUGE-W は語順に厳しく、それに対して ROUGE-S は語順の制約は比較的緩いことが特徴である。

また、編集距離の 3 つの操作である substitutions、insertions、deletions に対して、shift の操作を加えた 4 つの操作に基づく自動評価尺度として TER（Translation Error Rate）[7]がある。TER は以下の式により得られる。スコア値は 0.0 以上であり、値が小さいほど評価は高くなる。基本的には WER と同様の特徴を有するが、shift の操作が加わったことにより、WER に比べて人間の直観により近くなり、操作回数のカウントの仕方がよりクリアになっている。

$$\mathrm{TER} = \frac{substitutions + insertion + deletion + shift}{average\ \#\ of\ reference\ words}$$

その他には GTM（General Text Matcher）[8]、CDER（Cover Disjoint Error Rate）[9]、NMG_WN（Normalized Mean Grams-Word Number）[10]なども提案されている。NMG_WN は最大の gram 数に基づく構文的自然性に着目した自動評価尺度である。

6.2.5 WMT における自動評価尺度

2006 年以降に毎年開催されている統計翻訳のためのコンテスト型ワークショップ Workshop on Statistical Machine Translation では、2008 年から評価タスクが開催され、自動評価のためのメタ評価が実施されている。そこで、2009 年以降の WMT においてこれまでに上位にランキングされた自動評価尺度をいくつか取り上げる。

様々な自動評価手法を一つのセットとしてスコア値を決定する自動評価尺度 ULC（Uniform Linear Metric Combinations）[11]は 2009 年の WMT においてトップにランキングされている。スコア値は以下の式により得られる。Xが自動評価尺度のセットを表している。$x(a, R)$は MT 訳文 a と参照訳 R に対する自動評価尺度 x のスコア値である。したがって、このスコア値は個々の自動評価尺度のスコア値の平均を表している。具体的には、WER、PER[12]、TER、BLEU、NIST、GTM、ROUGE、METEOR が使用されている。更には、様々なレベルの語彙知識を自動評価尺度と組み合わせることでより精度の高い評価を実現している。

$$ULC_{X(a,R)} = \frac{1}{|X|}\sum_{x \in X} x(a, R)$$

また、語彙情報を用いた n-gram 一致率に基づく自動評価尺度 MaxSim（Maximum Similarity Metrics）[13]も上位にランキングされている。スコア値は以下の式より得られる。MaxSim はトークン、活用、品詞、WordNet などの語彙情報を用いた n-gram 一致率に基づいている。式 *sim-score* は式 $score_s$ の平均を表している。s は MT 訳文と参照訳の 1 つのペアを意味し、S はすべてのペアを示している。式 $score_s$ の n は n-gram を表し、実際には n を 1 から 3 まで変化させている。$F_{s,n}$ は F 値を表している。したがって、例えば n=2 の場合には、bigram における適合率と再現率を用いた F 値ということになる。

$$sim\text{-}score = \frac{1}{|S|}\sum_{s=1}^{|S|} score_s$$

$$score_s = \frac{1}{N}\sum_{n=1}^{N} F_{s,n}$$

その他には TE（Textual Entailment）に基づく RTE（Recognition of Textual Entailment）を用いた自動評価尺度[14]も上位にランキングされている。具体的には MT 訳文と参照訳の間で相互に含意関係であれば良い翻訳、それに対して、相互に含意関係ではない場合には悪い翻訳であるとする。これらの自動評価尺度は 2009 年の WMT において上位にランキングされている。ヨーロッパ言語 4 言語と英語間の翻訳による MT 訳文と人手評価、そして、参照訳と人手評価の 5 つのスピアマンの順位相関の平均は、システムレベルで ULC が 0.83、MaxSim が 0.80、RTE が 0.79 であった。また、文レベルでのケンドールの順位相関は ULC が 0.54、RTE が 0.53、MaxSim が 0.52 であった。

これまでに述べた自動評価尺度は静的な言語知識にどれほど依存するかによって 3 つに分類される[15]。まずは、言語知識を用いない自動評価尺度であり、BLEU、TER などが該当する。次いで、浅い言語知識に基づく自動評価尺度でこれは METEOR、MaxSim が該当する。そして、深い言語知識に基づく自動評価尺度でこれは RTE、ULC が該当する。WMT の上位にランキングされている自動評価尺度は基本的に言語知識に基づく自動評価尺度である。人手評価との高い相関を有する自動評価尺度を実現するためには、様々な静的な言語知識に頼らざるを得ない。しかし、その場合、言語知識の更新作業などのメンテナンスに労力を要し、多言語への適用が困難になるという問題を抱えることになる。

次に、2010 年以降の WMT で上位にランキングされている自動評価尺度について述べる。ヨーロッパ言語から英語への翻訳において、2010 年の WMT でトップにランキングされた自動評価尺度は SemPOS（Semantic Part-of-Speech）[16]であった。SemPOS は、内容語の重なりで評価する。その際には、単語を t-lemma と呼ばれる標準形に変換して比較を行う。2011 年の WMT でトップにランキングされた自動評価尺度は MTeRater-Plus[17]であった。MTeRater は人間が書いたエッセイの自動表採点用に考案された ETS e-rator®の素性に基づく機械学習を用いている。採点用の素性としては grammar、usage、mechanics、style、organization、development、lexical、complexity、vocabulary usage の誤り採点を用いている。更に、MTeRater-Plus は MTeRater の素性に BLEU、TERp[18]、METEOR などの自動評価尺度の素

性を加えている。この MTeRater-Plus では参照訳が不要である。そして、2012 年の WMT では SemPOS がトップにランキングされ、その次に AMBER（A Modified BLEU, Enhanced Ranking Metric）[19]がランキングされた。AMBER は n-gram の適合率と再現率の関数で定義されたスコア値に対して、いくつかのペナルティにより重み付けを行っている。SBP（Strict Brevity Penalty）は BLEU の BP が全訳文の長さの和のみを問題にし、個々の訳文を考慮していないという問題を解決するためのペナルティとなっている。SRP（Strict Redundancy Penalty）は訳文が長い場合に対するペナルティである。CKP（Chunk Penalty）は METEOR の *Pen* と同様にチャンクに基づくペナルティである。また、AMBER では正規化した順位相関係数（スピアマン、ケンドール）も導入している。そして、AMBER では多くのパラメータを用いているため、downhill simplex 法でチューニングを行っている。

英語からヨーロッパ言語の翻訳においては、2010 年の WMT で上位にランキングされた自動評価尺度の一つは TESLA（Translation Evaluation of Sentences with Linear-programming-based Analysis）[20]であった。TESLA は 2011 年の WMT でもトップにランキングされている。TELSA では、WordNet による同義語を用いている。また、MT 訳文と参照訳との間の n-gram 一致率において、n-gram に重みを与えられるように、n-gram のマッチングを線形計画法で解いているのが特徴である。2011 年の WMT でトップにランキングされた TESLA-M は n ユニグラムの F 値の平均を用い、TESLA-F は SVM-rank でトレーニングを行ったものである。2012 年の WMT で上位にランキングされた自動評価尺度 TerrorCat[21]や BlockErrCats[22]は翻訳誤りを分類し、誤りのカテゴリごとに翻訳品質に与える影響を調べることで評価を行う。その場合、言語によって重要な要素が異なる。英語では語順の誤りが重要であり、チェコ語やドイツ語では屈折の誤りが重要となる。

また、2011 年の WMT では統計翻訳のチューニングの観点より、自動評価尺度のメタ評価を行っている。その結果、BLEU が最も有効であることが明らかとなっている。更に 2012 年の WMT では参照訳を用いることなく、機械学習の技術に基づき自動評価を行うことを目的とした Quality Estimation Task が実施された。上位にランキングされた自動評価尺度は SDLLW_MP5bestDeltaAvg、SDLLW_SVM や UU_bltk、UU_best であった。SDLLW[22]では、ベースライン素性、デコーダ素性、独自素性の 3 種類の素性を用いて機械学習を行っている。ベースライン素性としては原言語・目的言語でのトークン数、言語モデルの確率、低頻度・高頻度の n-gram の割合、句読点の数など 17 の素性を用いている。デコーダ素性としては、Moses デコーダの内部コストから 8 つの素性を用いている。独自素性では、原言語側の未知語数など 20 の素性を用いている。UU_best[22]では、ベースライン素性 17 を含む 99 の素性と翻訳品質の推定に tree kernel を取り入れた SVR（Support Vector Regression）を利用している。UU_bltk ではベースライン素性 17 と tree kernel を取り入れた SVR を利用している。

6.2.6 WMT 以外の自動評価尺度

WMT 以外で発表されている自動評価尺度について述べる。PORT（Precision-Order-Recall Tunable metric）[23]は統計翻訳のチューニングに利用することを目的とした自動評価尺度であ

り、言語リソースを必要とせず、また、高速処理が可能である。PORT は以下の式より得られる。以下の式は Q_{mean} と新しい順位相関係数である ν の α 乗の調和平均を表している。Q_{mean} は n-gram 適合率の平均 P_a と n-gram 再現率の平均 R_a にパネルティ SBP と SRP をそれぞれかけたものの二乗平均である。SBP と SRP は自動評価尺度 AMBER で使用しているペナルティと同様である。

$$PORT = \frac{2}{1/Q_{mean}(N) + 1/\nu^{\alpha}}$$

$$Q_{mean}(N) = \sqrt{\frac{(P_a(N) \times SBP)^2 + (R_a(N) \times SRP)^2}{2}}$$

また、AM-FM（Adequacy-oriented component of the metric – Fluency-oriented component of the metric）[24]は参照訳が不要であり、CL-LSI（Cross-Language Latent Semantic Indexing）を利用することで原文と訳文のコサイン類似度を求める AM と n-gram 言語モデルによる確率である FM の重み付き調和平均を求めることで評価を行う。AM-FM のスコア値は以下の式より得られる。

$$AF\text{-}FM = \frac{AM \times FM}{\alpha AM + (1-\alpha)FM}$$

HyTER（Hybrid Translation Edit Rate）[25]は同義表現をアノテーション可能なツールを用いて、等価な表現の組み合わせにより膨大な参照訳の集合を表すことで評価を行う。その際には、MT 訳文と参照訳を人間が見て作成した参照訳を基準にして計算した TER である HTER（Human TER）[26]を利用する。また、距離計算は有限状態オートマトンにより効率的に計算している。

一方、意味論に基づく翻訳評価の研究が盛んに行われている。SRL（Semantic role labeling）のマッチングによる採点に基づき評価を行う MEANT[27]や RTE に基づく意味的テキスト類似度を用いて評価を行う Sagan[28]などはその一つである。

更に、近年、日本で提案された自動評価尺度である IMPACT（Intuitive common Parts ConTinuum）[29]と RIBES（Rank-based Intuitive Bilingual Evaluation Score）[30]について述べる。IMPACT は MT 訳文と参照訳間の共通部分を一意に決定し、その共通部分に基づき再現率と適合率との F 値をスコア値として出力する。共通部分の決定は以下の式により決定する。

$$pos_w = \left(1.0 - \left|\frac{posX(c)}{m} - \frac{posY(c)}{n}\right|\right)$$

$$RS = \left(\sum_{c \in LCS} \left(length(c)^{\beta} \times pos_w\right)\right)^{\frac{1}{\beta}}$$

式中の pos_w は共通部分の相対的な位置のずれを表している。そして、RS において共通部分の長さに対する重み付けとして pos_w を用い、この RS が最も大きい共通部分列を一意に選択する。次いで、決定された共通部分列を構成する共通部分の長さに基づき適合率 P と再現率 R を求め、それらの F 値をスコア値とする。P と R、そして、スコア値は以下の式より得られる。P と R の計算式において α、β はパラメータである。IMPACT では語順の異なる共通部分が出現する場合には、カウンタ i がインクリメントされるため、α（<1.0）が負の重み付けとしての役割を持つ。β は共通部分の長さに対する重み付けパラメータである。IMPACT は文単位の評価において人手評価との高い相関が得られる。

$$P = \left(\frac{\sum_{i=0}^{RN} \left(\alpha^i \sum_{c \in CC} length(c)^\beta \right)}{m^\beta} \right)^{\frac{1}{\beta}}$$

$$R = \left(\frac{\sum_{i=0}^{RN} \left(\alpha^i \sum_{c \in CC} length(c)^\beta \right)}{n^\beta} \right)^{\frac{1}{\beta}}$$

$$IMPACT = \frac{(1+\gamma^2)PR}{\gamma^2 P + R}$$

$$\gamma = \frac{P}{R}$$

RIBES は語順が大きく異なる言語対を対象に、訳語の違いよりも語順を重視することで人手評価との間で高い相関を得ることを目的としている。RIBES のスコア値は以下の式より得られる。

$$RIBES = \frac{\tau + 1}{2} \times P^\alpha$$

$$(0 \leq \alpha \leq 1)$$

P はユニグラム適合率を表している。τ はケンドールの順位相関係数を示している。$(\tau+1)/2$ により順位相関係数が[0,1]の値をとるように正規化している。

6.2.7 まとめ

これまでに様々な自動評価尺度が提案されてきた。そして、WMT における評価タスクなど、自動評価尺度のメタ評価が実施されるようになったことで、自動評価尺度の精度向上が着実に進んでいる。更に、使用目的や前提条件に独自性を持たせることでその存在意義を見出そうとする研究も進んでいる。例えば、統計翻訳のチューニングに特化した自動評価尺度や参照訳の

使用を前提としない自動評価尺度などはその例である。このように自動評価尺度は多様な広がりを見せており、今後は評価精度の向上を追求しながらも、利用者のニーズに沿った様々な自動評価尺度の研究が進んでいくと考えられる。

参考文献

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu(2002) “BLEU: a Method for Automatic Evaluation of Machine Translation,” Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311-318.

[2] Philipp Koehn(2010) “Statistical Machine Translation,” Cambridge University Press

[3] NIST(2002) “Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics”

[4] Alon Lavie and Abhaya Agarwal(2007) “Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments,” Proceedings of the Second Workshop on Statistical Machine Translation, pp. 228–231.

[5] Geroge Leusch, Nicola Ueffing and Hermann Ney(2003) “A Novel String-to-String Distance Measure With Applications to Machine Translation Evaluation,” Proc. of MT Summit IX, pp.240-247.

[6] Chin-Yew Lin and Franz Josef Och(2004) “Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics,” Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311-318.

[7] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla and John Makhoul(2006) “A Study of Translation Edit Rate with Targeted Human Annotation,” Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA), pp. 223-231.

[8] Joseph P. Turian, Luke Shen and I. Dan Melamed(2003) “Evaluation of Machine Translation and its Evaluation,” Proc. of MT Summit IX, pp.386-393.

[9] Gregor Leusch, Nicola Ueffing and Hermann Ney(2006) “CDER: Efficient MT Evaluation Using Block Movements,” Proceedings of EACL 2006, pp. 241-248.

[10] 江原暉将(2007) “新しい機械翻訳自動評価基準 NMG の提案”, Japio 2007 Year Book, pp.238 -241.

[11] Jesús Giménez and Lluís Márquez(2007) “Heterogeneous Automatic MT Evaluation Through Non-Parametric Metric Combinations,” Proceedings of IJCNLP, pp. 319-326.

[12] Keh-Yih Su Ming-Wen Wu and Jing-Shin Chang(1992) “A New Quantitative Quality Measure for Machine Translation Systems,” Proceedings of the 14th International Conference on Computational Linguistics, pp.433-439.

[13] Yee Seng Chan and Hwee Tou Ng(2008) “MAXSIM: An Automatic Metric for Machine Translation Evaluation Based on Maximum Similarity,” Proceedings of the Metrics-MATR Workshop of AMTA-2008, pp. 319-326.
[14] Sebastian Padó, Michel Galley, Dan Jurafsky, Christopher D. Manning(2009) “Textual Entailment Features for Machine Translation Evaluation,” Proceedings of the 4th Workshop on Statistical Machine Translation.
[15] Chang Liu, Daniel Dahlmeier and Hwee Tou Ng(2010) “TESLA: Translation Evaluation of Sentences with Linear-programming-based Analysis,” Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pp.354-359.
[16] Matouˇs Mach´aˇcek and Ondˇrej Bojar(2011) “Approximating a Deep-Syntactic Metric for MT Evaluation and Tuning,” Proceedings of the 6th Workshop on Statistical Machine Translation, pp92-98.
[17] Kristen Parton, Joel Tetreault, Nitin Madnani and Martin Chodorow(2011) “E-rating Machine Translation,” Proceedings of the 6th Workshop on Statistical Machine Translation, pp. 108–115.
[18] Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz(2009) "Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric", Proceedings of the Fourth Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the Association for Computational Linguistics.
[19] Boxing Chen and Roland Kuhn(2011) “AMBER: A Modified BLEU, Enhanced Ranking Metric,” Proceedings of the 6th Workshop on Statistical Machine Translation, pp. 71–77.
[20] Chang Liu, Daniel Dahlmeier and Hwee Tou Ng(2010) “TESLA: Translation Evaluation of Sentences with Linear-programming-based Analysis,” Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pp. 354–359.
[21] Mark Fishel, Rico Sennrich, Maja Popovi´c and Ondˇrej Bojar(2012) “TerrorCat: a Translation Error Categorization-based MT Quality Metric,” Proceedings of the Seventh Workshop on Statistical Machine Translation, pp. 64-70.
[22] Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut and Lucia Specia(2012) “Findings of the 2012Workshop on Statistical Machine Translation,” Proceedings of the 7th Workshop on Statistical Machine Translation, pp.10–51,
[23] Boxing Chen, Roland Kuhn and Samuel Larkin(2012) “PORT: a Precision-Order-Recall MT Evaluation Metric for Tuning,” Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp. 930–939.
[24] Rafael E. Banchs and Haizhou Li(2011) “AM-FM: A Semantic Framework for Translation Quality Assessment,” Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pp. 153–158.

[25] Markus Dreyer and Daniel Marcu(2012) “HyTER: Meaning-Equivalent Semantics for Translation Evaluation,” 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 162–171.
[26] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul(2006) “A Study of Translation Edit Rate with Targeted Human Annotation,” Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pp. 223-231.
[27] Chi-kiu Lo and Dekai Wu(2011) “MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames,” Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 220–229.
[28] Julio Castillo and Marina Cardenas(2012) “SAGAN: A Machine Translation Approach for Cross-Lingual Textual Entailment,” First Joint Conference on Lexical and Computational Semantics (*SEM), pp. 721–726.
[29] Hiroshi Echizen-ya and Kenji Araki(2007) “Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum,” Proceedings of the Eleventh Machine Translation Summit, pp.151-158.
[30] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, Hajime Tsukada(2010) “Automatic Evaluation of Translation Quality for Distant Language Pairs,” Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 944–952.

６．３　「空気の読める機械翻訳」の評価方法

(株)東芝　研究開発センター
知識メディアラボラトリー
鈴木　博和

6.3.1　背景

6.3.1.1　空気の読める機械翻訳

「空気が読める」というのは周りの状況や文脈、人間関係などを適切に判断し正しい行動を執ることができるということである。機械翻訳の場合、この「行動」とは即ち「翻訳する」ことを示す。従って「空気が読める翻訳」というのは「周りの状況や文脈、人間関係などを適切に判断した上でその場に合った正しい翻訳をする」ことを表す。

近年、スマートフォンなどでは音声翻訳アプリケーションに注目が集まっている。これらの音声翻訳技術と今までの文書翻訳技術との大きな違いは、前者が後者よりもより「コミュニケーション」に重心を置いたものであるという点である。従って「空気の読み具合」がユーザの使い心地に大きく影響を及ぼす。

もちろん、SNS の普及により文書・文章の翻訳においてもコミュニケーションを目的としたものは存在するが、音声翻訳に於いては、専門分野の論文やマニュアルなどをその対象とした従来の機械翻訳とは明らかに指向性が異なる。

従ってこのような技術の研究開発においては、その成果（即ち「空気の読み具合」性能）を正しく評価できる枠組みが必要である。これは従来型の機械翻訳評価方法とは一線を画すものであり、実現には相当な議論や困難も予想されるが、実現されれば機械翻訳研究者だけでなく、それを利用するエンドユーザ、機械翻訳の導入を検討している部門担当者、機械翻訳を翻訳のアシストツールと考えている実務翻訳者、などなど多くの人々にとって有益であると考えられる。

人手評価	・機械翻訳結果を見るのは結局「人」なので、システムの出力訳文の評価には最終的には人手が必要不可欠 ・評価者が指定された観点で訳文を評価し、指定された操作を行う（点数付け、ランキングなど） ・評価者の言語能力（ネイティブ・非ネイティブ）、文化的背景、経験などが評価に影響する ・主観的に評価するので評価者間での評価が一致しない(inter-consistency が低い) ・評価者自身の評価が一定でない (intra-consistency が低い) 場合もある ・評価には人・時間を要するため高コスト
自動評価	・「評価」→「改善」のサイクルを短くすることを目的に導入されるが、あくまでも人手評価を補完する(予測する)ための手法として認識されており、人手評価の必要性を排除するものではない

	・参照訳(正解)にどれくらい近いかを評価する。 ・原文に対して複数通りの翻訳が可能なので、参照訳は複数用意するのが好ましい。 ・用意した参照訳の質によって評価が変わる ・人手評価を良く予測できている手法(人手評価との相関が高い手法)がよい手法とされる ・低コスト

さらにこのような評価は、前述の専門分野の論文やマニュアルの翻訳を評価する時にも流用できるため、その意義は大きい。

しかし、現在のところこのような観点での機械翻訳評価方法は見当たらない。確認のため先ずは現状の機械翻訳評価方法についてまとめる。

6.3.1.2 機械翻訳の評価

機械翻訳技術の開発において訳文の「評価」は極めて重要であり、システムが出力した訳文をどのような観点でどう評価するか、については様々な手法が提案されている。

これら評価手法を大別すると**「人手評価」**と**「自動評価」**とに分類され、それぞれ主に以下のような特徴がある。

「評価」→「改善」のサイクルを短くし、迅速に機械翻訳技術を開発するためには、評価にかかるコストが低い自動評価手法の導入が必要である。一般的に、自動評価手法の良し悪しを判断するのに人手評価手法との相関が用いられるが、このとき人手評価には、評価者間・評価者内での評価のバラツキが少なく、高い信頼性を有することが求められる。しかし、人手評価手法は今までこのような観点ではあまり評価されておらず、そのため統一的に確立された枠組みが存在しない。

このような状態であるにもかかわらず、自動評価手法は確立された手法が存在しない人手評価手法との相関の高さで優劣を競っているのが現状である。人手評価手法が確立されていないので、それを精度の根拠にしている自動評価手法評価も疑問が残る。

そこで本報告書では、評価者間・評価者内での評価のバラツキが少なく信頼性の高い人手評価手法の方針検討を行った。この人手評価は主に訳語選択や文法の正しさのみを評価している従来の評価手法とは異なり、ニュアンスや状況による訳し分け、文化的背景なども視野に入れられるようになっており、未来の技術「空気の読める機械翻訳」を研究開発していく上での試金石ともなり得ると考える。

6.3.2 人手評価手法の調査

6.3.2.1 代表的な人手評価手法

まず、従来の代表的な人手評価手法とその特徴について述べる。

6.3.2.1.1 LDC-style

Advanced Research Projects Agency(ARPA)[1]で提案された **Adequacy(適確性)** と **Fluency(流暢性)** の 2 つの観点で評価を行う手法であり、LDC では以下の 5 段階で評価を行う[2]。

Rank	Adequacy(適確性)	Fluency(流暢性)
5	ALL source information	FLAWLESS target sentence
4	MOST source information	GOOD target sentence
3	MUCH source information	NON-NATIVE target sentence
2	LITTLE source information	DISFLUENT target sentence
1	NONE source information	INCOMPREHENSIBLE
0	NO TRANSLATION RESULT = IMPOSSIBLE TO JUDGE	

シンプルな評価方法であり、評価に不慣れな評価者も評価しやすいことから、人手評価手法として良く用いられる手法であるが、次のような問題点があることが指摘されている：

- Adequacy と Fluency は個別に評価できることを前提としているが、実際に評価を行うと、Adequacy と Fluency との間に強い相関がある[3]。
- 評価者が主観的に評価を行うが、点数が荒いので、評価者間の一致度(agreement)が低い[4]。即ち同じ訳文に対してある評価者は 4 点を付け、別の評価者は 2 点をつけるような場合がある。これは評価者の評価の厳しさや点数の付け方の解釈の違いによる。

6.3.2.1.2 Ranking

評価者間の評価の相関が高くなるように、訳文に対して点数をつけることはせず、ランキングの観点で評価を行う手法である。代表的なものとして 4th EACL Workshop on SMT, 2009 で提案された手法[5]：

- 複数システムの出力を提示する
- 各翻訳結果を Best から Worst までランキングする
- 原文や参照訳を見ずに翻訳結果を編集し、後で原文・参照訳ペアを提示し、acceptable かどうか判定する。

---

[1] White, J., O'Connell, T. and O'Mara, F. (1994) "The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches".*Proceedings of the 1st Conference of the Association for Machine Translation in the Americas. Columbia, MD* pp.193—205

[2] http://projects.ldc.upenn.edu/TIDES/Translation/TranAssessSpec.pdf

[3] Chae et al. “Predicting the fluency of text with shallow structural features; case studies of machine translation and human-written text”, ACL, 2009

[4] Callison-Burch et al. “Further Meta-Evaluation of Machine Translation”, 3rd workshop on SMT, 2008

[5] Callison-Burch, Philipp Koehn et al. “Finding of the 2009 Workshop on Statistical Machine Translation” Proceedings of the 4th EACL Workshop on SMT, 2009

や、2 システムのみを提示し、よい方のみを選択していく binary comparison ベースの手法[6]がある。後者の場合は 1 回あたりの評価時間は少なく済むが、総評価回数が多くなってしまうという問題点がある。

6.3.2.2 Industry quality initiative

研究において機械翻訳文の品質を評価するのとは別に、産業界においても翻訳品質を評価・保証するのは極めて重要であり、翻訳文書が顧客の要求を満たしているかどうか、必要な品質に達しているかどうかを判断する指標や指針が提案されている。ここでは産業界での翻訳品質評価の指標・指針などについて述べる。

6.3.2.2.1 American Translation Association(ATA) accreditation program

産業翻訳者の能力の certification program であり、特に MT に関して言及しているわけではないが、"Into-English Grading Standards"が定義されており、翻訳結果の英文をどのような観点で評価すればよいかの参考になる。Standards は評価の観点とあわせて例も示している。

Approximately synonymous terms and translations of cognates

A term that a grader considers less than ideal is acceptable if all five of the following conditions apply:

- one (not necessarily the first) of its definitions in the AH Dictionary is acceptable in context or the term used is listed as a synonym of an acceptable term (in that term's appropriate sense);
- it is not generally considered to be nonstandard English and is not labeled in the AH Dictionary as belonging to a register that is unacceptable in context;
- the usage is grammatically appropriate;
- the dominant sense of the word (or the word as part of a phrase) used by the candidate does not create ambiguity or distortion of meaning; and
- the use of a synonym does not prevent or impede understanding (for instance, in an idiom or set phrase). (See the "Idioms" and "Phrasal verbs" entries in these Standards for additional examples.)

Examples:

- *transport* used for *transportation* in the sense of *the act of being transported.* Acceptable.
- *nowadays* for *today* (meaning *these days*). Acceptable.
- *Nevertheless, throwing money at the problem did not provide a solution.* Preferred rendition.
- *However, throwing money at the problem did not provide a solution.* Acceptable (*nevertheless* appears as AH definition 4 of *however*).

But:

- *I was very tired but decided to go to the concert nevertheless.* Correct.
- *I was very tired but decided to go to the concert however.* Usage error.
- *His sister was also a doctor.* Correct.
- *His sister was too a doctor.* Usage error.
- *State control is an important factor in psychology.* Intended meaning is *control of mental states*; dominant meaning is *government control*. Ambiguity error.

ATA CERTIFICATION PROGRAM Into-English Grading Standards (version 2010)の例

6.3.2.2.2 EN-15038: European Quality Standard for Translation Service

翻訳サービスが翻訳者以外の人物による**「翻訳と校閲」**作業の 2 段階に分けて遂行されること

[6] David Vilar et al., "Human Evaluation of Machine Translation Through Binary System Comparisons", 2nd Workshop on SMT, 2007

を義務付け、翻訳者と校閲者の専門的適正を明確に定義づけしている。

6.3.2.2.3 Society of Automotive Engineers J2450

自動車サービスに関する文書の翻訳品質を保証･維持するために､人手で評価するための metric が定義されている。人手評価をシステマティックに行いスコアを算出して評価するため現実性・実用性が高い｡以下はこの J2450 のサンプルであり､始めに示してあるのがスコアシートである。このスコアシートでは各項目に重みと数を記入し、最終的に重みの合計値を単語数で割った値をスコアとして算出するようになっている。
次に示しているのが、各評価項目の概要である。

APPENDIX A

TRANSLATION METRIC SCORE SHEET

A.1 See Figure A1

SAE J2450
Translation Metric Score Sheet

Error Type	Num * Serious	Num * Minor	Category Weighted Score
Wrong Term Score WT	_____ * 5 _____ +	_____ * 2 _____ =	_____
Syntactic Error Score SE	_____ * 4 _____ +	_____ * 2 _____ =	_____
Omission Score OM	_____ * 4 _____ +	_____ * 2 _____ =	_____
Word Structure/Agreement Score SA	_____ * 4 _____ +	_____ * 2 _____ =	_____
Misspelling Score SP	_____ * 3 _____ +	_____ * 1 _____ =	_____
Punctuation Error Score PE	_____ * 2 _____ +	_____ * 1 _____ =	_____
Miscellaneous Error Score ME	_____ * 3 _____ +	_____ * 1 _____ =	_____

Document Score: (sum of weighted scores ÷ number of words in source language document)

_____ + _____ + _____ + _____ + _____ + _____ + _____ = _____ Sum of Weighted Scores

_____ Sum of Weighted Scores ÷ _____ Number of Words in Source Text =

_____ ***Overall Document Weighted Score***

Example Category Score in Syntactic Error category with major weight 4 and minor weight 2, assuming 3 major syntactic errors and 4 minor syntactic errors:

3 * 4 + 4 * 2 = 20

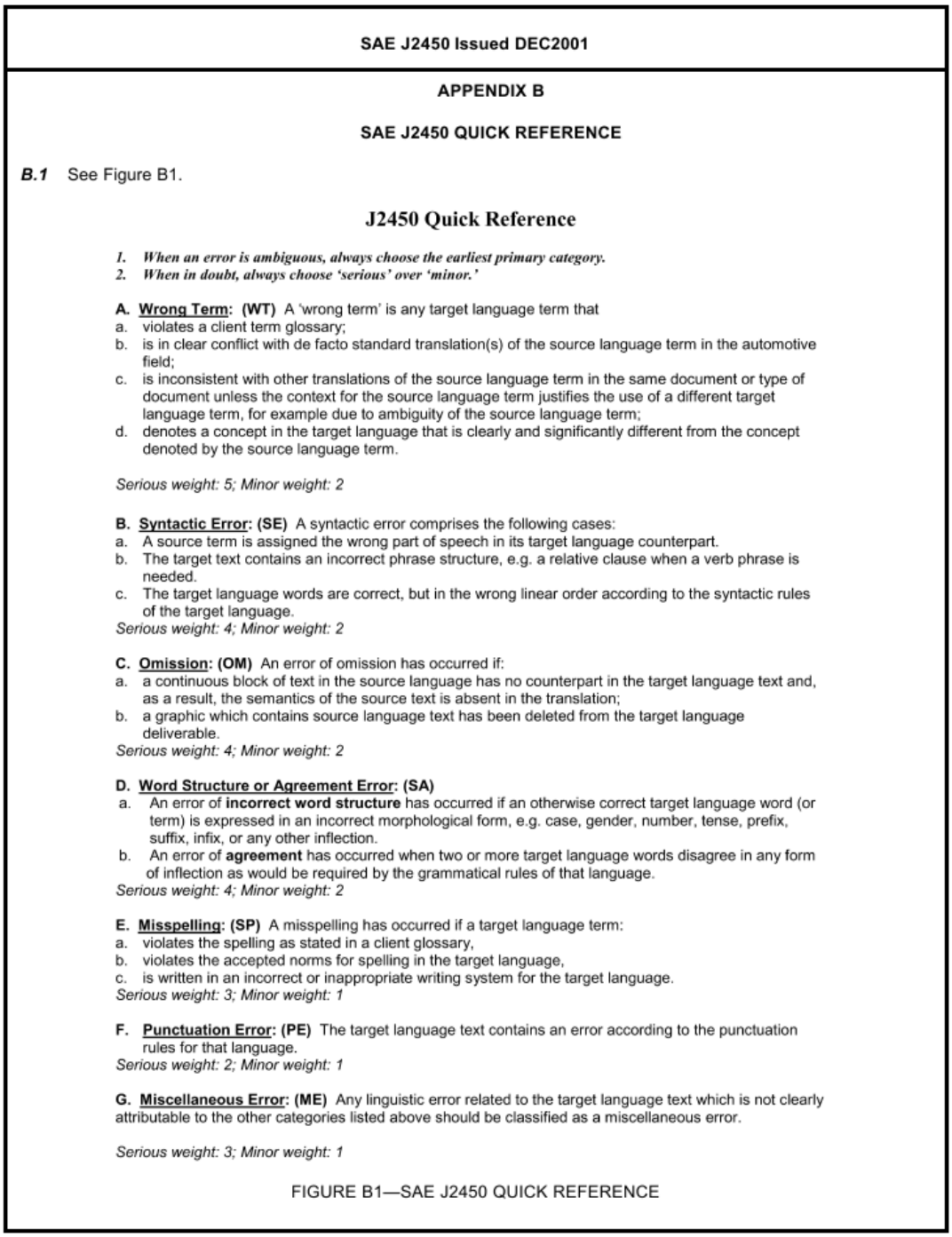
Example Document Score: (in a document with 330 source text words)

14 + 20 + 10 + 8 + 9 + 1 + 7 = 69 (Sum of Weighted Scores in all 7 categories)

69 (Sum of Weighted Scores) ÷ 330 (Number of Words in Source Text) =

0.209 ***(Overall Document Weighted Score)***

FIGURE A1—TRANSLATION METRIC SCORE SHEET



SAE J2450 Issued DEC2001

APPENDIX B

SAE J2450 QUICK REFERENCE

B.1 See Figure B1.

J2450 Quick Reference

1. ***When an error is ambiguous, always choose the earliest primary category.***
2. ***When in doubt, always choose 'serious' over 'minor.'***

A. Wrong Term: (WT) A 'wrong term' is any target language term that
a. violates a client term glossary;
b. is in clear conflict with de facto standard translation(s) of the source language term in the automotive field;
c. is inconsistent with other translations of the source language term in the same document or type of document unless the context for the source language term justifies the use of a different target language term, for example due to ambiguity of the source language term;
d. denotes a concept in the target language that is clearly and significantly different from the concept denoted by the source language term.

Serious weight: 5; Minor weight: 2

B. Syntactic Error: (SE) A syntactic error comprises the following cases:
a. A source term is assigned the wrong part of speech in its target language counterpart.
b. The target text contains an incorrect phrase structure, e.g. a relative clause when a verb phrase is needed.
c. The target language words are correct, but in the wrong linear order according to the syntactic rules of the target language.

Serious weight: 4; Minor weight: 2

C. Omission: (OM) An error of omission has occurred if:
a. a continuous block of text in the source language has no counterpart in the target language text and, as a result, the semantics of the source text is absent in the translation;
b. a graphic which contains source language text has been deleted from the target language deliverable.

Serious weight: 4; Minor weight: 2

D. Word Structure or Agreement Error: (SA)
a. An error of **incorrect word structure** has occurred if an otherwise correct target language word (or term) is expressed in an incorrect morphological form, e.g. case, gender, number, tense, prefix, suffix, infix, or any other inflection.
b. An error of **agreement** has occurred when two or more target language words disagree in any form of inflection as would be required by the grammatical rules of that language.

Serious weight: 4; Minor weight: 2

E. Misspelling: (SP) A misspelling has occurred if a target language term:
a. violates the spelling as stated in a client glossary,
b. violates the accepted norms for spelling in the target language,
c. is written in an incorrect or inappropriate writing system for the target language.

Serious weight: 3; Minor weight: 1

F. Punctuation Error: (PE) The target language text contains an error according to the punctuation rules for that language.

Serious weight: 2; Minor weight: 1

G. Miscellaneous Error: (ME) Any linguistic error related to the target language text which is not clearly attributable to the other categories listed above should be classified as a miscellaneous error.

Serious weight: 3; Minor weight: 1

FIGURE B1—SAE J2450 QUICK REFERENCE

SAE J2450

6.3.2.3 CEFR

英語への翻訳を考えた際、訳文の品質を評価するには英文の品質を評価する必要がある。産業翻訳のようにプロの翻訳家が翻訳したものではなく、様々なレベル（非ネイティブ～ネイティブ）の人が作成した英文の品質を評価する手法も調査した。その中で心理学的・統計学的見地から手法が構築され、確立されているのは教育学の分野である。そこで、Cambridge ESOL(English for

Speakers of Other Language)の試験が準拠している CEFR(Common European Framework of Reference)に着目した。

今回提案する人手評価手法は、この CEFR の基本概念をベースに構築している。従って、ここではこの CEFR について簡単に説明する。

6.3.2.3.1 絶対評価と相対評価

以前、英語教育においては成績の付け方は相対的につけられていた。これは試験を行い、その成績上位者の何名を 5 にする、といった成績の付け方である。このような相対評価はクラスによって差を生むことになる。即ち同じ成績 5 でも、レベルが高いクラスと低いクラスとでは英語の実力に差が生じていることになる。

この問題を鑑み、近年では絶対評価(到達度評価)を行うように変わってきている。これは

- 英語学習の指導要綱を設定
- 指導要綱に基づいたレベルの試験を作成
- 指導要綱に基づきその試験の何点以上を合格とするか(合格点)を設定する
- 受験した生徒の点数が合格点以上であれば必要な英語の到達度に達していると判断する(合格)。

このような絶対評価を行うと、今までとは異なりクラス全員が 5(全員が所定の学習レベルに到達)という状況や全員が不合格(全員が所定の学習レベルに未達)という状況もあり得ることになる。

機械翻訳の評価においてもこのような絶対評価手法の確立が重要と考える。この手法で評価すると、たとえば A 社機械翻訳システムはネイティブ小学生レベル、B 社機械翻訳システムはネイティブ大学生レベル、といった評価スコアとその意味するレベルとの対応が明らかにできるだけでなく、実際に「高品質な」機械翻訳システムと判断するためには、評価スコアで何点を採らなければならないか、といった目標値も設定することができる。この目標値は他システムと比較して設定できるだけでなく、その設定値の意味自体も非常に説得力のある値となる。

CEFR は絶対評価手法であるので、今後はこれをベースにすることにした。

6.3.2.3.2 Criteria と Standard

絶対評価では criteria(評価規準)と standard(基準)を設定する必要がある。criteria は前述の指導要綱や英語のレベルに相当する。standard は criteria に対応するスコアの境界を表わす。

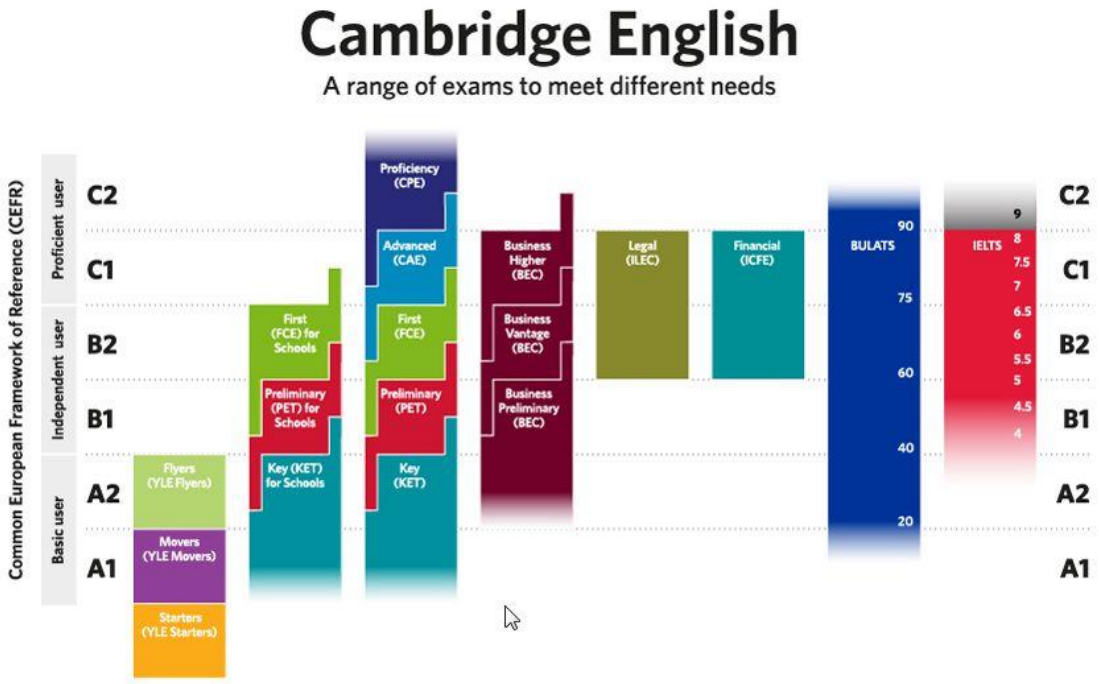
CEFR で設定している criteria には以下のようなものがある：

Table C4: WRITTEN ASSESSMENT CRITERIA GRID

	Overall	Range	Coherence	Accuracy	Description	Argument
C2	Can write clear, *highly accurate and* smoothly flowing complex texts in an appropriate and effective *personal* style *conveying finer shades of meaning. Can use* a logical structure which helps the reader to find significant points.	Shows great flexibility in *formulating* ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.	Can create coherent and cohesive texts making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices.	Maintains consistent *and highly accurate* grammatical control of *even the most complex language forms. Errors are rare and concern rarely used forms.*	Can write clear, smoothly flowing and fully engrossing stories and descriptions of experience in a style appropriate to the genre adopted.	Can produce clear, smoothly flowing, complex reports, articles and essays which present a case or give critical appreciation of proposals or literary works. Can provide an appropriate and effective logical structure which helps the reader to find significant points.
C1	Can write clear, well-structured *and mostly accurate* texts of complex subjects. *Can underline* the relevant salient issues, *expand and support* points of view at some length with subsidiary points, reasons and relevant examples, and *round* off with an appropriate conclusion.	Has a good command of a broad range of language allowing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say. *The flexibility in style and tone is somewhat limited.*	Can produce clear, smoothly flowing, well-structured text, showing controlled use of organisational patterns, connectors and cohesive devices.	Consistently maintains a high degree of grammatical accuracy; *occasional errors in grammar, collocations and idioms.*	Can write clear, detailed, well-structured and developed descriptions and imaginative texts in a mostly assured, personal, natural style appropriate to the reader in mind.	Can write clear, well-structured expositions of complex subjects, underlining the relevant salient issues. Can expand and support point of view with some subsidiary points, reasons and examples.
B2	Can write clear, detailed *official and semi-official* texts on a variety of subjects related to his field of interest, synthesising and evaluating information and arguments from a number of sources. *Can make a distinction between formal and informal language with occasional less appropriate expressions.*	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, using some complex sentence forms to do so. *Language lacks, however, expressiveness and idiomaticity and use of more complex forms is still stereotypic.*	Can use a number of cohesive devices to link his/her sentences into clear, coherent text, though there may be some "jumpiness" in a longer text.	Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstandings.	Can write clear, detailed descriptions of real or imaginary events and experiences marking the relationship between ideas in clear connected text, and following established conventions of the genre concerned. Can write clear, detailed descriptions on a variety of subjects related to his/her field of interest. Can write a review of a film, book or play.	Can write an essay or report that develops an argument systematically with appropriate highlighting of some significant points and relevant supporting detail. Can evaluate different ideas or solutions to a problem. Can write an essay or report which develops an argument, giving some reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options. Can synthesise information and arguments from a number of sources.
B1	Can write straightforward connected texts on a range of familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence. *The texts are understandable but occasional unclear expressions and/or inconsistencies may cause a break-up in reading.*	Has enough language to get by, with sufficient vocabulary to express him/herself with some circumlocutions on topics such as family, hobbies and interests, work, travel, and current events.	Can link a series of shorter discrete elements into a connected, linear text.	Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more *common* situations. *Occasionally makes errors that the reader usually can interpret correctly on the basis of the context.*	Can write accounts of experiences, describing feelings and reactions in simple connected text. Can write a description of an event, a recent trip – real or imagined. Can narrate a story. Can write straightforward, detailed descriptions on a range of familiar subjects within his field of interest.	Can write short, simple essays on topics of interest. Can summarise, report and give his/her opinion about accumulated factual information on a familiar routine and non-routine matters, within his field with some confidence. Can write very brief reports to a standard conventionalised format, which pass on routine factual information and state reasons for actions.
A2	Can write a series of simple phrases and sentences linked with simple connectors like "and", "but" and "because". *Longer texts may contain expressions and show coherence problems which makes the text hard to understand.*	Uses basic sentence patterns with memorized phrases, groups of a few words and formulae in order to communicate limited information mainly in everyday situations.	Can link groups of words with simple connectors like "and", "but" and "because".	Uses simple structures correctly, but still systematically makes basic mistakes. *Errors may sometimes cause misunderstandings.*	Can write very short, basic descriptions of events, past activities and personal experiences Can write short simple imaginary biographies and simple poems about people.	
A1	Can write simple isolated phrases and sentences. *Longer texts contain expressions and show coherence problems which make the text very hard or impossible to understand.*	Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.	Can link words or groups of words with very basic linear connectors like "and" and "then".	Shows only limited control of a few simple grammatical structures and sentence patterns in a memorized repertoire. *Errors may cause misunderstandings.*	Can write simple phrases and sentences about themselves and imaginary people, where they live and what they do, etc.	

表の最左列にある A1～C1 は試験受験者の英語レベルを表わしており、A1 が最も低く、C2 が最も高くなっている。上記 criteria は各英語レベルの者が有している実力を表わしている。従って、たとえばクラス C2 と判断された者は、上記 criteria の C2 の行に対応した英語の実力があると判断できる。

参考までに、Cambridge ESOL の種類と対応する CEFR クラスとを図示すると以下のようになる：



（出典:http://www.cambridgeesol.org/about/standards/cefr.html）

例えば、Cambridge PET に合格すると、B1(Threshold)の実力があると判断されることになる。

このように、criteria と standard を設定することにより絶対評価を行うことが可能となる。しかし、問題はこれらをどのように設定するかである。特に機械翻訳の場合は特に問題や課題を設け、それを受験させるわけではないので設定が難しい。以降では機械翻訳の場合における人手評価手法について述べる。

6.3.3　機械翻訳の人手評価手法の提案

以降では CEFR の概念を継承しながら、機械翻訳向けにアレンジを行い、人手評価手法の提案を行う。

6.3.3.1　Criteria Setting

まずは CEFR の criteria グリッドを参考にし、ランクは A1/A2/B1/B2C1/C2 の 6 段階を設けることとする。次に考慮すべきは各ランクに対応した機械翻訳の品質(Translation Quality, TQ)で

ある。実は TQ を設定するためには、「良い翻訳とはどのような翻訳か？」が明らかでなければならない。そのためにはそもそも「翻訳とは何か？」について考える必要がある。

6.3.3.1.1 「翻訳」の定義

「翻訳」をどのように定義するかについては様々な考え方が提案されている。大まかに二分すると **Equivalent Effect** という考え方と **Translation Loss** という考え方がある。これらについて簡単に説明する。

6.3.3.1.1.1 Equivalent Effect

「翻訳」とは目的言語話者が訳文を見たときに、原言語話者が原文を見たときと等価な(equivalent)効果(effect)が得られるような訳文を生成することであると考えられる。これが Equivalent Effect の基本的な概念である。これを提唱したのが Nida(1964)[7]であり、2 種類の equivalence、**formal equivalence**(内容と同様に形式にもフォーカス)と **dynamic equivalence**(言語的特徴や文化を考慮し、読み手の effect にフォーカス)を提案し多大な影響を与えた。

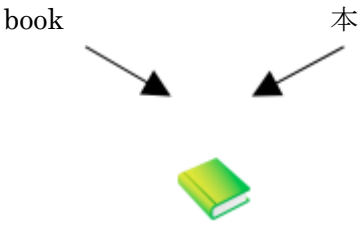
最も広く用いられている equivalence モデルは Koller(1979)[8]の 5 タイプ equivalence と Baker(1992)[9]の 5 タイプ equivalence である。

Koller は以下の 5 つの equivalence を定義している：

- **Denotative Equivalence**

context-independent な概念で原言語と目的言語とで同一の対象を指すような equivalence を表わす。

（例）



- **Connotative Equivalence**

両言語間で同一の communicative value を持つ equivalence を表わし、以下のような例がある。

（例）

- Speech level：elevated/poetic/normal/colloquial/slang/vulgar
- Socially determined usage：student/military/aristocratic/….
- Geographical relation or region：American English/Australian English/….
- Pompous：“I told you a million times!”

---

[7] Nida, Eugene (1964): “Toward a Science of Translation”, Leiden: Bril

[8] Werner Koller (1992), Einführung in die Übersetzungswissenschaft, Heidelberg / Wiesbaden 1979 (first edition)

[9] Baker, M(1992): In Other Words. A Coursebook on Translation: London, Routledge.

- Euphemistic

Euphemistic	
Toilet	Lavatory/ restroom
Toilet paper	Bathroom tissue
Baggage collectors	Sanitation workers
Die	Pass away/ kick the bucket/go to the heaven/ breathe your last breath/ be gone
Poor	Unable to make ends meet
Fat	Overweight/ chubby
Handicapped	Disabled/ physically challenged
Homeless	Without a roof over one's head

- Common/Uncommon

Old English	Modern English
Should auld acquaintance be forgot? And never brought to mind Should auld acquaintance be forgot? And days of Auld Lang Syne	Should we forget our old friends? And not remember our old friends? What happens if we forget people we used to know, And forget the past?
For Auld Lang Syne, my dear For Auld Lang Syne We'll take a cup of kindness yet For Auld Lang Syne	To remember the good times, In order to not forget the good old days Let's have another drink, And remember the good old days.

- **Text-normative Equivalence**
  原文と訳文とで同じフォーマットで記述されるものを表わす。
  (例) Sincerely yours = 敬具
- **Pragmatic/Dynamic Equivalence**
  読み手に書き手と同じ効果をもたらす equivalence を表わし、**communicative equivalence** ともいう。
- Formal Equivalence
  詩や歌詞のように韻や比ゆ、リズムなどの equivalence を表わし、**expressive equivalence** ともいう

一方 Baker の 5 タイプの equivalence は各言語レベルで equivalence をとらえた以下のようなものである：

- Equivalence at word-level
- Equivalence above word-level
- Grammatical Equivalence
- Textual Equivalence
- Pragmatic Equivalence

これらの equivalence の概念は「良い翻訳」を定義する上で重要である。例えば Connotative

Equivalence を考慮すれば、訳文の fluency/naturalness を大きく向上させることができると考えられる。

ただし、equivalent effect を持っているかどうかを判断することは難しく、基本的には読み手の反応・感想から判断するのが一般的のようだ。従って訳文の equivalence を工学的に数値化するのは困難だが、人手評価の観点としては重要と考える。

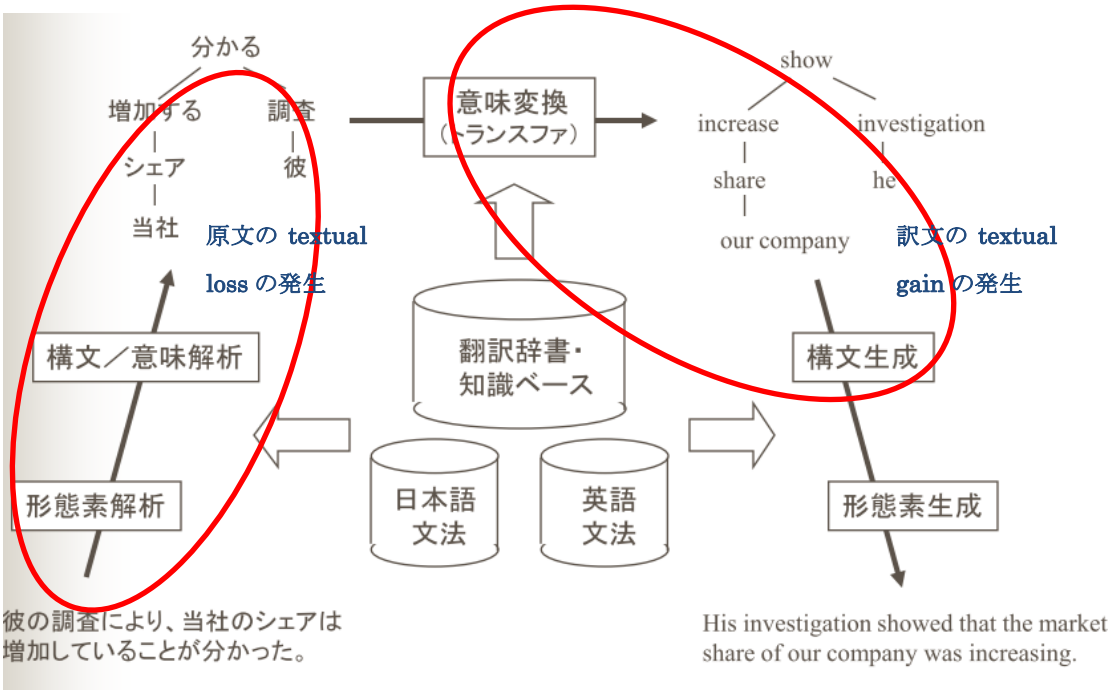
6.3.3.1.1.2 Translation Loss

Hervey and Higgins(1992)[10]は前述の equivalence の概念に疑問を持った。彼らによれば翻訳は必ずある程度の**意味的損失(loss)**を含むものであり、翻訳においては完全に等価な訳を作ろうとするのではなく、この loss を如何に抑えるかが重要であるとした。

Translation Loss には原文の textual feature の loss と訳文にしか存在しない textual feature の付加による gain の 2 種類があるとされている。

我々の機械翻訳は解析―変換―生成からなるルールベースシステムであるが、原文の解析段階で原文に存在する textual loss が派生し、変換～生成の段階で textual gain が発生していると考えれば、各段階でこれらの loss を抑えることが高品質の訳文を出力することにつながると解釈できる。

しかし、equivalence のときと同様に loss を工学的にどのように判断するかは非常に難しい。これらを扱うことができれば、ルールベース機械翻訳システムの各段階ごとに品質を向上させることが可能となる。



[10] Harvey, Sandor; Ian Higgins (1992): “Thinking Translation. A Course in Translation Method: French to English”, London: Routledge

6.3.3.1.2 Criteria Grid of MT Quality

以上を考慮し定義した criteria grid を示す：

Level		Qualitative Factors				
		Vocabulary Range & Vocabulary Control	Grammatical Accuracy	Socio-Linguistic Appropriateness	Coherence & Cohesion	Fluency & Flexibility
Proficient User Level	C 2	can convey finer shades of meaning precisely	maintains consistent and highly accurate grammatical control	can express fully the socio-linguistic and socio-cultural implications	can create coherent and cohesive texts making full and appropriate use of a variety of organizational patterns	uses expressions with natural, smooth flow (so smooth that the reader is hardly aware of MT)
	C 1	can translate clearly in an appropriate style on a wide range	consistently maintains a high degree of grammatical accuracy (occasional errors in grammar, collocations and idioms)	can use translation flexibly and effectively for social purposes	can produce clear, smoothly flowing, well-structured test	uses expressions with natural, smooth flow
Independent User Level	B 2	has a sufficient range of language to be able to give clear descriptions	high degree of grammatical control (does not make errors which cause misunderstandings)	can avoid crass errors of formulation	can use a number of cohesive devices to link sentences into coherent text (including minor jumpiness)	uses occasional less appropriate expressions
	B 1	has enough language to get by, with sufficient vocabulary	uses reasonably accurately a repertoire of frequently used patterns	-	can link a series of shorter discrete elements into a connected, linear text	the texts are understandable but occasional unclear expressions
Basic User Level	A 2	uses basic sentence pattern	errors may sometimes cause misunderstandings	-	can link groups of words with simple connectors	texts contain expressions which makes the text hard to understand
	A 1	has a very basic repertoire of words and simple phrases	limited control of a few simple grammatical structures	-	can link words or groups of words with very basic linear connectors (e.g. and/then)	text contain expressions which make the text very hard or impossible to understand
Incomprehensible Level	0	impossible to judge	non-grammatical	impossible to judge	impossible to judge	impossible to understand

品質評価の観点 qualitative factor として

- Vocabulary range & Vocabulary control
- Grammatical accuracy
- Socio-Linguistic appropriateness
- Coherence & Cohesion
- Fluency & Flexibility

の 5 つをあげた。Socio-Linguistic appropriateness は Koller の connotative equivalence を反映して導入している。

ランクは A1/A2/B1/B2/C1/C2 のほかに、NIST-style の 5 段階評価の rank 0 に相当する「判断不可」というランクを追加した。

通常、このような criteria grid を作成するには複数人の専門家がパネラーとして参加し、充分なディスカッションを重ねた上で決定される。従って最終的には言語学的・工学的側面から複数の専門家からのレビューが必要である。

6.3.3.2 Item Setting

次に TQ(Translation Quality)を評価するための項目、質問事項を決定する。これには

- 質問内容
- 何段階評価(or 何点満点の評価)にするか
- 質問数をいくつにするか

などを決める必要がある。質問は先に定義した criteria grid の各 qualitative factor を評価できる内容となっていなければならない。

Rui Rothe-Neves[11]は実際に翻訳品質を評価する質問事項を決め、それで評価者に評価させた場合、評価の相関や信頼度はどのくらいかを調査した。実際に、Nevas が設定した質問項目は以下のようなものであった：

Table 1 – Questions presented in assessment scale

1.	Does the text read fluently?
2.	Is the translation grammatically correct?
3.	Is the spelling correct?
4.	Are there unjustified inferences?
5.	Is the vocabulary adequate?
6.	Is the vocabulary used consistently throughout the text?
7.	Is the translation performed according to the assignment?
8.	Does the layout correspond to normal standards?
9.	Could the translation be used according to the style norms for this kind of text?
10.	Is the overall result satisfactory?

これらを 5 段階評価(1=Not at all, 2=A bit, 3=Somewhat, 4=Much, 5=Completely)で評価させている。加えて上記質問項目が妥当であるかどうかを判断させ、妥当であれば 1、妥当でなければ 0 を乗じることにより質問の内容の信頼性を検証している。

機械翻訳の品質評価の場合も、複数の専門家の意見を反映しながら、上記と同様に複数の質問内容と得点配分や grading を決定する必要がある。次に実際に評価を行い、評価者間の評価の相関係数や質問の信頼度を検証し、それらの項目が妥当かどうかを判定する必要がある。

本報告書では質問内容を決定するまでは至っていないが、それらが決定した後にどのように妥当性を検証していくかの手順について述べる。

検証は次の 3 つの観点で行う。ここでは 7 つの質問項目(各 10 点満点)を設定して 6 人の評価者に評価させた場合の結果が以下の表のようになったと仮定して説明する。

[11] Rui Rothe-Neves, Universidade Federal de Minas Gerais, “Translation Quality Assessment for Research Purposes: An Empirical Approach”

		Item						
		Q1	Q2	Q3	Q4	Q5	Q6	Q7
Judge	J1	9	8	9	7	8	10	9
	J2	8	7	7	7	8	8	8
	J3	8	7	8	6	6	8	7
	J4	10	8	10	7	8	8	8
	J5	10	9	9	8	8	9	10
	J6	9	9	10	8	7	7	7
	Avg.	9	8	8.83	7.17	7.5	8.33	8.17

6.3.3.2.1 Concordance (Statistics for agreement)

評価者間の評価のバラツキが大きい場合、質問が適切でない可能性が高い。従ってどのくらい評価が一致しているかを調べる必要がある。このような場合に良く用いられるのはノンパラメトリック手法の **Kendall's W (Kendall's coefficient of concordance)** である。

項目 i に対する評価者 j によるランクを $r_{i,j}$ とすると、項目 i のトータルランクは

$$R_i = \sum_{j=1}^{m} r_{i,j},$$

で表わされる(m:評価者数)。n を項目数とすると、トータルランクの平均値は

$$\bar{R} = \frac{1}{2}m(n+1).$$

となる。

偏差２乗和 S は

$$S = \sum_{i=1}^{n} (R_i - \bar{R})^2,$$

のように定義される。このとき Kendall's W は以下の式で定義される[12]：

$$W = \frac{12S}{m^2(n^3 - n)}.$$

W は 0 から 1 の値をとり、1 に近いほど意見が一致していることを表わす。ちなみに上記表の例の場合は

$$W = 0.4815$$

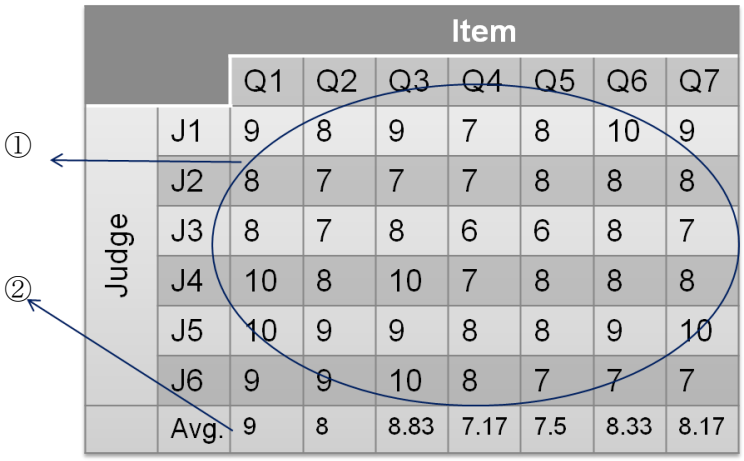
となる。

---

[12] Kendall, M. G.; Babington Smith, B. (Sep 1939). "The Problem of *m* Rankings". *The Annals of Mathematical Statistics* **10**(3): 275–287.

6.3.3.2.2 Consistency

評価者間の相関(intra-class correlation)によって評価者間の評価の一貫性を検証する。



		Item						
		Q1	Q2	Q3	Q4	Q5	Q6	Q7
Judge	J1	9	8	9	7	8	10	9
	J2	8	7	7	7	8	8	8
	J3	8	7	8	6	6	8	7
	J4	10	8	10	7	8	8	8
	J5	10	9	9	8	8	9	10
	J6	9	9	10	8	7	7	7
	Avg.	9	8	8.83	7.17	7.5	8.33	8.17

intra-class correlation は

上記②の分散/上記①の分散

で得ることができる。上の表でこれを計算すると 0.2962 となった。これを意味するものは全体の分散の約 70.4%(100－29.6)は評価者の評価の違いによって生じていることを表わしている。

6.3.3.2.3 Reliability

質問内容の信頼性は Cronbach's alpha[13]によって検証する。アンケート調査などで，対象とする領域のある特性を測定するために複数の質問項目への回答の合計値（特に尺度得点と呼ばれる）を使うことがある。Cronbach's alpha は尺度に含まれる個々の質問項目が内的整合性を持つかどうか（目的とする特性を測定する質問項目群であるか）を判定するために用いられる。

Cronbach's alpha は項目 j の不偏分散を S_j^2、k 個の変数の合計点の不偏分散を S_Y^2 としたとき，

$$\alpha = \frac{k}{k-1}\left(1 - \sum_{j=1}^{k} S_j^2 / S_Y^2\right)$$

であらわされる。上記の表で算出すると

$$\alpha = 0.83$$

となる。一般的にはαは 0.7 以上が好ましいとされているのでこの例の質問セットは内部整合性が高いといえる。

その他、質問数によって信頼性がどの程度変化するかを予想する指標として Spearman-Brown formula[14]がある。この指標などを参考にして最適な質問数の予測ができる。

---

[13] Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297-334.

[14] Spearman, Charles, C. (1910). Correlation calculated from faulty data. British Journal of Psychology, 3, 271-295.
Brown, W. (1910). Some experimental results in the correlation of mental abilities. British

6.3.3.3 Standard Setting

評価項目と得点配分が決定すると、criteria grid のランク(0, A1/A2/B1/B2/C1/C2)とスコアとを対応づける必要がある。即ち、例えば 100 点満点で何点以上を取れば C2 とみなせるのか、何点以下だったら A1 と判断できるのか、を決める作業である。

このような standard 決定手法として **body of work method** と **logistic regression** を挙げる。

6.3.3.3.1 Body of Work Method

あらかじめ様々な機械翻訳結果についてスコアをつけ、それらをスコア順にソートしておく。次にそれら機械翻訳文とスコアとを評価者に提示し、その訳文とスコアを見てシステムの能力を分類した場合、どのランクに入れるのが妥当かを判断してもらう。

下に示した表("Table 6.6: Summary of the Rangefinding Round")はスコア 13~54 について 15 人の評価者に分類させた例を表わす。例えばスコア 13 であれば 15 人全員が A1 に分類している。

ここで注目すべきは、同じスコアに対して、下位ランクへ分類した人数と上位ランクへ分類した人数が逆転している個所である。例えば、スコア 23 では A1 に分類した人が 10 人、A2 に分類した人が 5 人であるのに対し、次のスコア 24 を見ると、A1 分類が 7 人、A2 分類が 8 人となり、上位ランクの A2 に分類した人数が A1 に分類した人数を逆転している。

Journal of Psychology, 3, 296-322.

Folder	Score	A1	A2	B1	B2	Total
1	13	15	0			15
	15	15	0			15
	16	14	1			15
2	18	13	2			15
	19	11	4			15
	21	9	6			15
3	23	10	5			15
	24	7	8			15
	26	5	10			15
4	27	3	10	2		15
	28	0	12	3		15
	30	1	11	3		15
5	32		9	6		15
	33		11	4		15
	34		8	7		15
6	35		7	8		15
	36		8	7		15
	37		6	8	1	15
7	39		3	12	0	15
	41		1	14	0	15
	42		1	12	2	15
8	43			10	5	15
	45			11	4	15
	46			8	7	15
9	48			4	11	15
	49			1	14	15
	51				15	15
10	52				15	15
	53				15	15
	54				15	15

このように人数が逆転しているスコア近辺を「standard の候補」として認識することができる。

6.3.3.3.2 Logistic Regression

standard をピンポイントで決定するために使用する手法には**logistic regression(ロジスティック回帰分析)**がある。

ここでは因子は独立変数とし、線形モデルとして logistic regression を適用する。応答は standard に達している(1)か達していない(0)の 2 値で十分なので、logit を採用する。従って

$$\ln\frac{p}{1-p} = a + b * score$$

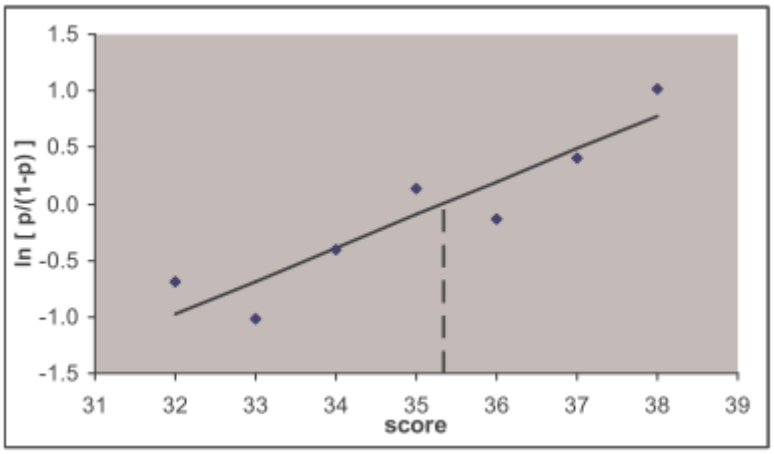
とモデル化できる、p は standard に到達する確率を表わす。

Score	A2	B1	B2	*p*	ln[*p*/(1-*p*)]
32	10	5		0.333	-0.6931
33	11	4		0.267	-1.0116
34	9	6		0.400	-0.4055
35	7	8		0.533	0.1335
36	8	7		0.467	-0.1335
37	6	9		0.600	0.4055
38	4	10	1	0.733	1.0116

上記のデータを基に logistic 回帰分析を行うと、係数 a、b が求まり、

$$a = -10.3744, \quad b = 0.29358$$

となる。これらに基づいてプロットすると以下のようになる：



cut-off score は standard に達したときの確率が 0.5 の場合に設定すればよいので

$$\ln\frac{0.5}{1-0.5} = 0 = a + b * \text{score}$$

従って、

$$\text{cut off score} = \frac{-a}{b} = \frac{10.3744}{0.29358} = 35.34$$

standard の cut-off スコアを 35.34 と予測することができる。

6.3.3.4 Agreement/Consistency

上記の Body of work method では評価者が機械翻訳結果を A1 から C2 に分類することになるが、その時の一致度(Agreement)と一貫性(Consistency)を検証する必要がある。agreement や consistency が低い場合は評価にバラツキがあることを意味するので、評価対象の機械翻訳文が良くない(判断が分かれるような訳文)可能性がある。

6.3.3.4.1 Agreement

２評価者間の評価の agreement を調べるために良く用いられる指標は Cohen's Kappa[15]と呼ば

[15] Jacob Cohen(1960)：Educational and Psychological Measurement

れる統計量であり以下の式であらわされる：

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)},$$

Pr(a)は観測された評価者間の相対的 agreement の確率、Pr(e)は偶発的 agreement の仮定確率(観測された値を使って評価者が無作為に分類したときの確率を計算)を表わす。

κ の値は 0 から 1 の値をとり、以下のような解釈がなされる：

κ	解釈
< 0	一致無し
0.0 - 0.20	わずかな一致
0.21 - 0.40	まずまずの一致
0.41 - 0.60	十分な一致
0.61 - 0.80	相当な一致
0.81 - 1.00	ほとんど完全な一致

Cohen's Kappa は 2 評価者間の agreement を評価するが、複数評価者間の agreement を評価する場合は **Fleiss' Kappa** を用いる。Fleiss' Kappa は以下の式で与えられる：

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

ただし、

$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} P_i$$

$$= \frac{1}{Nn(n-1)}(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^2 - Nn)$$

$$\bar{P}_e = \sum_{j=1}^{k} p_j^2$$

である。Nは評価対象数（評価する機械翻訳文数）を表わし、n は対象ごとの評価数（評価者数）を表わす。

k は割り当てるカテゴリ数（0/A1/A2/B1/B2/C1/C2 の 7 段階なので 7）を表わしている。n_{ij} は評価対象 i に対してカテゴリ j を割り当てている評価者数を表わしている。

p_i は評価対象 i に対して一致している評価者の確率を表わす。

p_j は評価がカテゴリ j である割合を表し、以下の式であらわされる：

$$p_j = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij}, \qquad 1 = \frac{1}{n}\sum_{j=1}^{k} n_{ij}$$

従って上記 κ の計算式の分母$1 - \bar{P}_e$は「１－偶然の一致率」を表わし、分子$\bar{P} - \bar{P}_e$は「評価の一致率－偶然の一致率」を表わしている。

6.3.3.4.2 Consistency

評価者間の評価の一貫性(consistency)を評価するための指標としては相関係数が良く用いられる。統計の分野で相関係数として使われるのは Pearson の積率相関係数があるがこの手法は偏差の正規分布を仮定するパラメトリック手法であり、機械翻訳評価のようなケースでは正しい相関が得られない可能性がある。従って、このような仮定を置かないノンパラメトリックな方法として多く用いられる **Spearman の順位相関係数**を採用する。

Spearman の順位相関係数 ρ は：

$$\rho = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

で求められる。ただしここで

D= 対応する X と Y の値の順位の差

N= 値のペアの数

である。

例えば、下記のような結果になった場合を考える：

		Judge 2				
		A1	**A2**	**B1**	**B2**	**Total**
Judge 1	**A1**	7	2	1	1	11
	A2	1	10	2	1	14
	B1	1	2	12	2	17
	B2	0	1	0	7	8
	Total	9	15	15	11	50

この場合の Judge1 と Judge2 との間の Speaman の順位相関係数を求めると

$$\rho = 0.6731$$

となり両者には強い相関があるといえる。

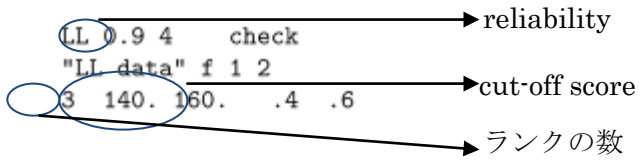
6.3.3.5 Validation

以上によって standard を決定すると、スコアと criteria grid のランク(0/A1/A2/B1/B2/C1/C2)とが対応づけられる。ここでは、上記のように決定した standard がどのくらいの accuracy・consistency を以て分類できるかを検証する。このような目的で使用する手法に **Livington & Lewis**[16]の手法がある。ツールとしては、BB-CLASS[17]を利用する。

---

[16] Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy

以下ではこのツールを使った例を示す。下記は BB-CLASS への入力データを表わす。




```
LL 0.9 4    check
"LL data" f 1 2
3  140. 160.    .4  .6

121 3      141 5      161 14     181 8
122 5      142 20     162 17     182 3
123 8      143 11     163 17     183 9
124 5      144 14     164 23     184 0
125 3      145 15     165 29     185 7
126 9      146 21     166 19     186 5
127 2      147 13     167 16     187 0
128 2      148 12     168 33     188 2
129 9      149 10     169 12     189 1
130 18     150 18     170 34     190 1
131 10     151 18     171 16
132 11     152 17     172 21
133 13     153 8      173 17
134 12     154 21     174 32
135 10     155 6      175 0
136 11     156 33     176 32
137 16     157 32     177 22
138 11     158 7      178 14
139 16     159 17     179 8
140 15     160 36     180 25
```

reliability は前述の Cronbach's alpha などを使って算出する。この例では 140 点と 160 点が cut-off score になっており、140 点未満・140 点以上 160 点未満・160 点以上の 3 つのランクに分類することを指示している(最初の 3 行)。その後は実際のスコアと頻度(ここでは人数)を表わしている。BB-CLASS 実行結果は以下のようになる：

```
**************************************************************************
*** BB-CLASS:   Beta-Binomial Classification Consistency and Accuracy ***
***                          Version 1.1                               ***
***                                                                    ***
***                       Robert L. Brennan                            ***
***                            CASMA                                     ***
***                       University of Iowa                          ***
***                                                                    ***
***                         December 2004                                ***
***                                                                    ***
***                       All Rights Reserved                          ***
```

of classifications based on test scores. Journal of Educational Measurement, 32, 179–197.

[17] Center for Advanced Studies in Measurement and Assessment. Univ. of Iowa

```
**************************************************************************

*** Livingston and Lewis Results ***

*** Listing of Control Cards in ccLL ***

LL 0.9 4       check
"LL data" f 1 2
3    140. 160. .4 .6

***********************************************

                  Number of examinees =       1000.00000
                    Input reliability =          0.90000
                 Effective test length =         49.70252
Effective test length (rounded) =          50
                              Lord's k =          0.00000     (binomial error model)

Number of Categories = 3

Cut Scores:      xcut[] = 140.00000 160.00000
              xprimecut[] =    21.91011    33.14607
                    tcut[] =     0.40000     0.60000

Category proportions in original data:
                     0.21400     0.31300     0.47300

***Moments used by Livingston and Lewis Procedure***

                    Mean           S.D.           Skew           Kurt            Min            Max

       x   155.244000     17.921006     -0.499583      2.543831   101.000000   190.000000
       p     0.609483      0.201360     -0.499583      2.543831     0.000000     1.000000
Raw=x'    30.474157     10.067981     -0.499583      2.543831     0.000000    50.000000

***Parameter Estimates for Beta Distribution***
```

alpha	beta	low limit	upp limit
2.666934	1.302899	0.000000	0.907239

Number of moments fit: 3

Moments (Raw, Fitted Raw, True)

	Mean	S.D.	Skew	Kurt
Raw	30.474157	10.067981	-0.499583	2.543831
Fitted Raw	30.474157	10.067981	-0.499583	2.515916
True	30.474157	9.554546	-0.546516	2.561804

Likelihood Ratio Chi-Square = 158.33372 (with df = 48)
Pearson Chi-Square = 151.27795 (with df = 48)
for cells with fitted frequencies greater than 0.00

Reliability (from above moments) = 0.90061
SEM (from above moments) = 3.17410

Numerical Integration

Numerical integration performed using 1000 equally-spaced quadrature points
and the true-score density for each interval

Sum of bivariate probabilities = 1.00000

*****ACCURACY RELATIVE TO EXPECTED OBSERVED SCORES GIVEN MODEL*****

	x0	**x1**	**x2**	**marg**
t0	**0.14951**	**0.01090**	**0.00000**	**0.16042**
t1	**0.06173**	**0.20016**	**0.01146**	**0.27335**
t2	**0.00045**	**0.12324**	**0.44255**	**0.56624**
marg	**0.21169**	**0.33430**	**0.45401**	**1.00000**

probability of correct classification = 0.79222
false positive rate = 0.02236; false negative rate = 0.18542

*****CONSISTENCY USING EXPECTED OBSERVED SCORES GIVEN MODEL*****

	x0	x1	x2	marg
x0	0.16625	0.04479	0.00066	0.21169
x1	0.04479	0.22121	0.06830	0.33430
x2	0.00066	0.06830	0.38505	0.45401
marg	0.21169	0.33430	0.45401	1.00000

pc = 0.77251; pchance = 0.36269; kappa = 0.64304
probability of misclassification = 0.22749

*****ACCURACY RELATIVE TO ACTUAL OBSERVED SCORES*****

	x0	x1	x2	marg
t0	0.15114	0.01021	0.00000	0.16135
t1	0.06240	0.18741	0.01194	0.26174
t2	0.00046	0.11539	0.46106	0.57690
marg	0.21400	0.31300	0.47300	1.00000

probability of correct classification = 0.79961
false positive rate = 0.02215; false negative rate = 0.17824

*****CONSISTENCY USING EXPECTED (row) VS. ACTUAL (column) OBSERVED SCORES*****

	x0	x1	x2	marg
x0	0.16806	0.04193	0.00068	0.21068
x1	0.04527	0.20712	0.07116	0.32355
x2	0.00066	0.06395	0.40116	0.46577

marg 0.21400 0.31300 0.47300 1.00000

pc = 0.77634; pchance = 0.36667; kappa = 0.64685
probability of misclassification = 0.22366

上記結果の太字の部分が検証結果を表わす。

もし、分布が既知(あるいは正確に予測されている)で、cut-off スコアが与えられれば、

- 各ランクにどのくらいの人数が分布するか決定できる → 真のランク
- モデルの仮定と reliability から、テストスコアに基づいてどれくらいの人数が各ランクに割り当てられるかが決定できる →予測ランク

以上を考慮し accuracy と consistency を算出している。

```
***ACCURACY RELATIVE TO EXPECTED OBSERVED SCORES GIVEN MODEL***

           x0       x1       x2      marg
 t0   0.14951   0.01090   0.00000   0.16042
 t1   0.06173   0.20016   0.01146   0.27335
 t2   0.00045   0.12324   0.44255   0.56624

marg  0.21169   0.33430   0.45401   1.00000

probability of correct classification = 0.79222
false positive rate = 0.02236;  false negative rate =  0.18542
```

上のマトリックスの行は真のランク、列は予測ランクを表わす。次のマトリックス

```
***CONSISTENCY USING EXPECTED OBSERVED SCORES GIVEN MODEL***

           x0       x1       x2      marg
 x0   0.16625   0.04479   0.00066   0.21169
 x1   0.04479   0.22121   0.06830   0.33430
 x2   0.00066   0.06830   0.38505   0.45401

marg  0.21169   0.33430   0.45401   1.00000

pc = 0.77251;  pchance = 0.36269;  kappa = 0.64304
probability of misclassification = 0.22749
```

は、列・行とも予測ランクを表わしている。3 つ目のマトリックス

```
***ACCURACY RELATIVE TO ACTUAL OBSERVED SCORES***

           x0       x1       x2      marg
 t0   0.15114   0.01021   0.00000   0.16135
 t1   0.06240   0.18741   0.01194   0.26174
 t2   0.00046   0.11539   0.46106   0.57690

marg  0.21400   0.31300   0.47300   1.00000

probability of correct classification = 0.79961
false positive rate = 0.02215;  false negative rate =  0.17824
```

では、行が真のランク、列が実データでのランクを表わしている。

```
***CONSISTENCY USING EXPECTED (row) VS. ACTUAL (column) OBSERVED SCORES***

           x0        x1        x2      marg
  x0   0.16806   0.04193   0.00068   0.21068
  x1   0.04527   0.20712   0.07116   0.32355
  x2   0.00066   0.06395   0.40116   0.46577

marg   0.21400   0.31300   0.47300   1.00000

pc = 0.77634;  pchance = 0.36667;  kappa = 0.64685
probability of misclassification = 0.22366
```

最後のマトリックスでは、行が予測ランク、列が実データでのランクを表わしている。
上記の probability of correct classification の値や kappa の値が指標になる。

こうして validation までチェックすることにより、今までの人手評価で問題となっていた評価のバラツキや一貫性の問題が起こらない、高品質の人手評価が可能になる。

6.3.4 まとめ

本報告書では「空気の読める翻訳」の評価方法に関し、特に人手評価方法の構築について、具体的に考察したことをまとめた。これらはまだまだ荒削りであり実際にこれを構築・応用するには多くの専門家（機械翻訳研究者、実務翻訳者、言語学者など）の意見も加味しなければならない。

今後は AAMT/Japio 特許翻訳研究会での活動を通して、特許調査に特化した機械翻訳の人手評価手法の確立を目指す。その際、本報告書内で検討したことが少しでも参考になれば幸いである。

６．４　NTCIR-9，NTCIR-10 特許機械翻訳タスクでの人手評価

(独)情報通信研究機構　後藤 功雄

6.4.1　はじめに

筆者らは特許機械翻訳の評価を行っている．本稿では，過去に主催した評価型ワークショップ NTCIR-9 特許機械翻訳タスク[1]（2010 年 7 月～2011 年 12 月），および現在実施中（2012 年 1 月～2013 年 6 月）の NTCIR-10 特許機械翻訳タスク[1]での人手評価について説明する．人手評価について説明する前に，まず特許機械翻訳タスクの概要を説明する．

6.4.2　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．

6.4.3　人手評価の必要性

訳質の評価方法の１つとして，自動評価がある．多くの自動評価は参照訳と翻訳結果とを比較して参照訳に類似する翻訳結果の訳質が高いとみなす．自動評価は評価コストが低く短時間で結果が分かるために便利である．翻訳手法が似ているシステム間の比較では信頼性が高い．そのため，システムのパラメータ調整や，手法が似ているシステム間の比較に有用である．

しかし，自動評価は完璧ではない．ルールベース機械翻訳システムと統計機械翻訳システムなど手法が大きく異なるシステム間の比較は信頼性が低い[2,1]．また，自動評価値からはどれだけの訳質が達成できたかが明確ではない． 例えば，代表的な自動評価である BLEU の値が例えば 20, 30, 40, 50 それぞれの場合において，翻訳元文の意味が分かる文の割合は不明である．

このように自動評価は完全ではないため，最も信頼性が高い評価は現時点では人手評価であるといえる．特許機械翻訳タスクでは，人手評価をメインの評価として実施している．

[1] http://ntcir.nii.ac.jp/PatentMT-2/

6.4.4 NTCIR-9 で実施した人手評価

本節では，NTCIR-9 で実施した人手評価について説明する．

6.4.4.1 NTCIR-9 での評価の設計

機械翻訳の用途を「情報収集のための利用」に設定した．このことから，次の条件において評価することにした．（1）後編集を前提としない．（2）翻訳結果から原文の内容を理解したい．この用途のための評価基準として，Adequacy と Acceptability という 2 つの評価基準を用いた．Adequacy の主な目的は，システム間の比較である．Acceptability の主な目的は，原文の意味が理解できる文数の割合を明らかにすることである．以下，Adequacy と Acceptability の評価基準について述べる．

6.4.4.2 Adequacy

NTCIR-9 で用いた adequacy は，評価は文単位で，翻訳結果における翻訳元文中の意味の保持度合を 5 段階で評価する．表 1 に adequacy の評価基準を示す．この 5 段階の None～All のグレード設定は文献[3]に基づく．この評価基準は，評価レンジが広く，低い訳質から高い訳質まで分類できるという特徴がある．また，客観的な基準ではないため，訳質の低いシステムばかりあるいは高いシステムばかりを評価する場合は，それらの訳質の差が明らかになるように 5 分類するという運用が可能である．すなわち実質的に相対評価としての運用が可能であると考えられる．

表 1：Adequacy 評価基準

5	All
4	Most
3	Much
2	Little
1	None

6.4.4.3 Acceptability

Acceptability は，文の訳質評価のために新たに設計した基準である．Acceptability の設計方針は次の通りである．

・各評価グレードに意味を持たせる．例えば，「原文の意味が理解できる」や「原文の意味が理解できて文法が正しい」などを評価グレードとして用いる．

・流暢さも同時に評価する．原文の意味が理解できる場合に限り，流暢さも評価する．これによって，原文の内容を反映しているかどうかを無視した評価を避けられる．

Acceptability は，評価は文単位で，文レベルの意味を評価する．以下に Acceptability の評価基準を示す．この評価基準による評価結果からは，何割の訳文で原文の意味が伝わるかを知ることができる．

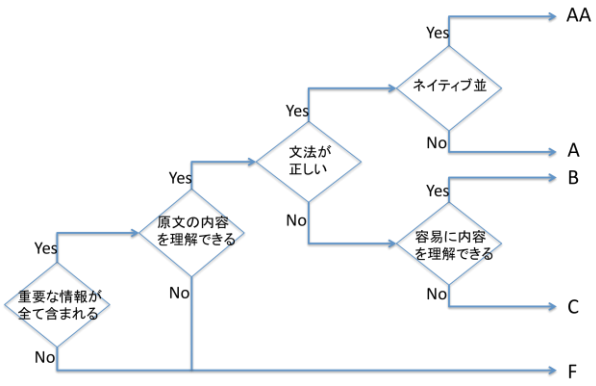


図 1：Acceptability 評価基準

6.4.4.4　各評価基準の課題

Adequacy の課題

基準に客観性がなく，評価の各グレードが実際にどのような訳質であるか明確でない．例えば，原文の意味が理解できる文の割合は不明である．また，湧き出し語に対する定義がない．目的言語側での表現の流暢さに対する評価である Fluency は別評価とされているが，Adequacy と Fluency の役割分担が明確でない．

Fluency の課題

Adequacy とペアで評価に用いられることが多い Fluency の課題についても説明する．Fluency は流暢さのみを評価するため，原文の意味を反映していない流暢な訳文に高いスコアを与える．しかし，原文の意味を反映している訳文のみに流暢さのスコアを与えなければ，訳質の評価にならない．例えば，入力の日本語が「今日は晴れです．」で，出力の英語が「Hello!」の場合，これは誤訳で訳質は明らかに低いが，目的言語側での流暢さだけを考えれば全く問題点がないため，Fluency は最高点である．このような誤訳に高いスコアを付与しても翻訳の評価として適切でないことは明らかである．

Acceptability の課題

中／低品質な訳文に対する評価の解像度が低い．例えば，一部誤訳でも全部誤訳でも最低評価になってしまう．そのため，現状の機械翻訳の精度では，システム間の比較は Adequacy の方が向いていると考えられる．

また，実際の情報収集の用途での有用性とのギャップがある．部分的にしか正しく訳せていない文でも，重要な部分が分かれば有用である場合がある．しかし，Acceptability は，部分的にしか正しく訳せていない文は最低評価になってしまう．なぜこの文を翻訳する必要があるのかという目的を明確にしなければ，部分的に正しく訳せている文の有用性を評価することは困難である．

6.4.4.5　評価に必要な文数の検討

本節では，信頼出来る評価を得るために必要な文数について検討する．もしテストデータが非常に少量（例えば 1 文のみ）であれば，たまたまその文だけうまく翻訳できた，またはできなかったという可能性が高く，システムの比較評価の信頼性は高いとはいえない．信頼性がある評価にするためには，ある程度の文数の訳質を評価する必要がある．

NTCIR-9 での評価は次のように実施した. 各システムあたりランダムに選択した 300 文を評価した．評価者数は 3 人（有償）である．各評価者は各システムあたり 100 文を評価した．原文が同じである翻訳結果は同時に評価した.

この検討のため，評価結果をシステム毎に前半 150 文と後半 150 文に分けて比較を行い，評価データが 150 文の場合で，adequacy によるシステム間の比較がどの程度一致するかを調べた．システム間の比較には，正規化のために，Adequacy の値ではなくシステム間の相対評価値（詳細は文献[1]の pairwise score を参照）に換算して比較を行った．日英翻訳の結果を図 2 に示す．Half-1 は前半の 150 文の評価結果，Half-2 は後半 150 文の評価結果，All は 300 文の評価結果を示している．また，表 2 に日英，英日，中英での異なる 150 文でのシステム比較の Pearson 相関係数を示す．この相関係数は，Half-1 と Half-2 のシステムの相対評価値の相関を示している．

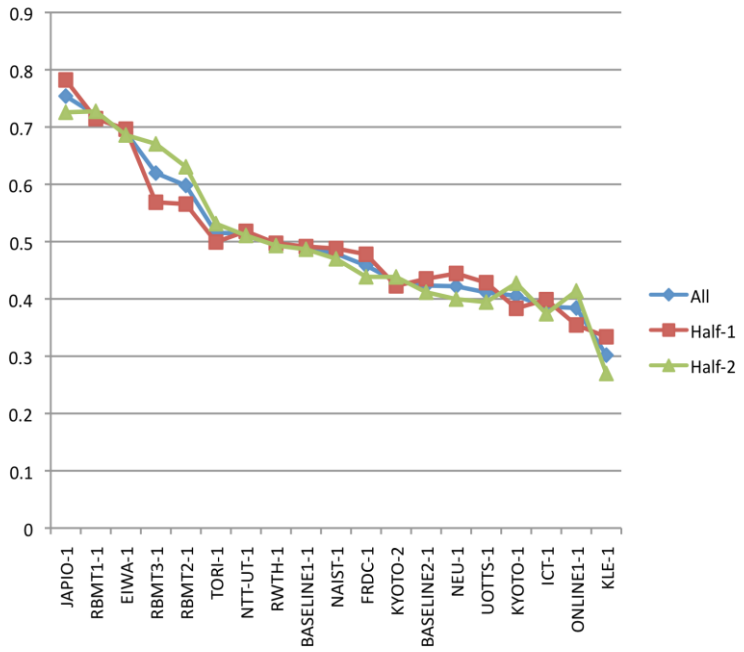


図 2：異なる 150 文でのシステム比較（日英翻訳）

表 2：異なる 150 文でのシステム比較の Pearson 相関係数

	Pearson 相関係数
日英	0.940
英日	0.985
中英	0.963

異なる 150 文でのシステム間の比較結果から次のことが分かった．図 2 より，Half-1 と Half-2 の間では，多少のスコアの違いはあるが上位のシステムと下位のシステムが入れ替わるような大きな違いはない．また，表 2 より，日英以外に，英日，中英の翻訳結果でも異なる 150 文のテストデータによる評価結果において，システム間の比較で高い相関を示している．このことから，システム間の評価の差が大きい場合，150 文でのシステム間の比較結果の信頼性は高いといえる．なお，300 文の場合は文数が多いためそれ以上の信頼性があると考えられる．

6.4.5 NTCIR-10 での特許審査評価

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，特許審査での有用性に基づく評価（特許審査評価）を NTCIR-10 特許機械翻訳タスクで実施することにした．本節では，特許審査評価について紹介する．

6.4.6 評価の考え方

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献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．

6.4.6.1 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．

6.4.6.2　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する．

（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 3 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 3 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 3 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 3 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．

表 3：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。

6.4.6.3　評価基準

特許審査評価の評価基準を表 4 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う．引用文書単位で評価する．

表 4：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）

6.4.7 まとめ

NTCIR-9 で実施した人手評価および NTCIR-10 で新たに実施中の人手評価について述べた．人手評価は，信頼性が高い評価方法であるが，時間や手間のコストは大きい．そのため，一定の信頼性を確保した上で評価するデータ量は少なくしたい．NTCIR-9 の評価結果から，ランダムに選択した 150 文の評価でもシステム間の差が大きければ信頼性は高いことが分かった．

NTCIR-9 で実施した文の訳質の人手評価基準は次の特徴がある．Adequacy は，システム比較に有用である．しかし，各グレードの訳質は明確ではない．Acceptability は，原文の意味が理解できる文数の割合が分かる．

NTCIR-10 で実施中の特許審査評価は，特許審査という応用における有用性の評価であり，なぜこの文を翻訳する必要があるのかという目的を明確にすることで，部分的に訳せている文に対しても有用性を評価することができる．特許審査評価の評価結果は 2013 年 6 の NTCIR-10 Workshop にて発表される予定である．

参考文献

[1] Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K. Tsou. 2011. Overview of the patent machine translation task at the NTCIR-9 workshop. In Proceedings of NTCIR-9, pages 559–578.

[2] Chris Callison-Burch, Miles Osborne, Philip Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics, pages 249–256.

[3] LDC. 2005. Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Translations Revision 1.5. Linguistic Data Consortium. http://projects.ldc.upenn.edu/TIDES/Translation/TransAssess04.pdf.

６．５　テストセット評価について

（株）富士通研究所　長瀬 友樹

6.5.1　はじめに

機械翻訳の研究開発者にとって翻訳文の品質評価は不可欠なものであり、これまでにも多くの評価方法が提案されてきた。実際に用いられている評価方法としては、正確性や流暢性の観点で人間が評価する方法と BLEU[1]に代表される自動評価法に大別されるが、どちらの方法も評価結果は文単位あるいはドキュメントの単位で品質が数値化されるものである。
本稿では、あるエンジンの翻訳結果が他と比べてどこが優れて（劣って）いるのかという種類の定性的な情報を得ることができる「テストセット評価」に着目し、テストセット評価と他の評価方法との比較、テストセット評価の課題と今後の展望等について述べる。

6.5.2　テストセット評価とは

本稿においてテストセット評価とは、翻訳言語対毎に整備した原文-正解訳文の対と、それぞれの文対に付与した設問からなるデータの集合を指す。図１は JEIDA（日本電子工業振興協会）によって開発された英日翻訳評価用テストセットの一例である。[2]

```
【番号】   2.1.1.2-1   (= 【ID No.】)
【例文】   The trash can was thrown away.   (= 【Example】)
【訳文】   ごみカンは捨てられた。   (= 【Translation】)
【質問】   "can" が「カン/缶」のように名詞として訳されていますか？
           (= 【Q.】 Is "can" translated as a noun?)
【訳出例】 ○ (くず缶/ごみ容器/くず入れ)は(廃棄された/[投げ]捨てられた)。
           × ごみは捨てられ得る。
           (= 【Translation Samples】 literally meaning:
                 yes: The (garbage can/trash bin/litter bin) was (discarded/[thrown] dumped).
                 no: The trash can be discarded. )
【関連文】 The last will was opened.「最後の遺言書は開けられた。」
           (= 【Related Examples】 and the Japanese translation)
【参照項目】2.1.1.2-2, 2.1.1.2-3   (= 【Reference Items】)
【解説】   "can was" の並びから、"can" が助動詞でないことがわかる。
           (= 【Explanation】 The word order of "can was" shows that "can" is not an auxiliary
                 verb.)
```

図１　JEIDA のテストセットサンプル

図１の「例文」と「訳文」のラベルが付いたところが原文（英語）と参照訳（日本語）であり、「質問」のところにこの対訳例文で評価すべきポイントが設問形式で書かれている。この例では、

「”can”が「カン/缶」のように名詞で訳されていますか？」という設問なので、翻訳結果が「くずの缶は捨てられた」なら“Yes”、「ゴミは捨てられることができる」なら“No”と評価されることになる。

このように、テストセット評価とは例文の翻訳結果を各対訳例文に付与された設問に照らして”Yes”または”No”の評価を行うものである。したがって、テストセット評価の結果としては、例文数に対して”Yes”評価と”No”評価がそれぞれ何文あったかという定量評価と、どのような文で”No”が多かったかといった定性評価が同時に行えるという特徴がある。

6.5.3　テストセット整備の現状

図 1 で示した JEIDA の英日翻訳用テストセットは 1993 年に開発され、翌年には日英翻訳用テストセットが開発された。その後 JEIDA のテストセットを引き継ぐ形で、内元らによる日英翻訳用テストセット中の設問の自動設定、自動評価についての研究報告がある。[3)]

日中翻訳用テストセットに関しては、やはり JEIDA のテストセットをベースにして、AAMT（アジア太平洋機械翻訳協会）の課題調査委員会によって 2009 年に第一版が完成した。当初、AAMT の日中翻訳用テストセットは人間が一文ずつ”Yes”、”No”の評価を行うことを想定して開発されたが、2011 年に人間の代わりにプログラムが自動的に設問回答できる「設問ベース自動評価」が可能なテストセットへと発展を遂げた。[4)]

図 2 は人間評価と従来型自動評価との比較においてテストセット評価（人間設問評価と設問ベース自動評価）の位置づけを示したものである。コスト的観点ではテストセット評価は人間評価と従来自動評価の中間に位置づけられる。設問ベース自動評価は設問作成の分だけ従来自動評価に比べてコストがかかるが、文法項目ごとの評価結果が得られるというメリットがある。一方、人間設問評価と従来型人間評価との比較では、設問回答の所要時間（評価者負担）と評価結果の客観性確保の観点から設問評価にメリットがある。

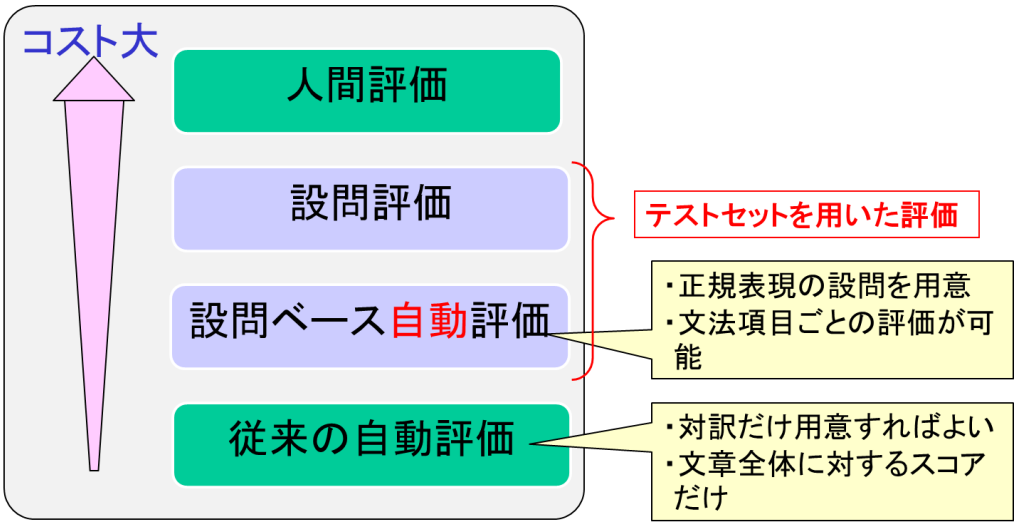


図 2　テストセット評価の位置づけ

ここで AAMT の設問ベース自動評価がどのように実現されたかについて触れておきたい。テストセットの設問を簡略化し、表１のように訳文の中に特定の文字列が含まれるか否かを問う内容としている。これによって、図３のように設問を正規表現ベースのプログラムに置きかえることを可能とし、設問評価の自動化が実現されている。

表１　日中テストセットの設問（抜粋）

Original Questions	Expanded Questions
「病気で」が「因为生病」になってますか？	「病気で」が、「因为生病」または「由于生病」または「因为毛病」または「由于毛病」または「因为病」または「由于病」または「因病」になっていますか？
「木で」が「用木头」になってますか？	「木で」が、「用木头」または「拿木头」になっていますか？

```
read a line;
if the line match /严|严厉|严格|厉害/ then
      print "Yes";
else
      print "No";
endif

read a line;
if the line match /去买东西|去购物|去逛街|去逛商店/ then
      print "Yes";
else
      print "No";
endif
```

図３　日中テストセットの設問評価プログラム（pseudo code）

6.5.4　テストセットによるＭＴ評価の事例

6.5.4.1　評価実験の方法

前節で述べた人間設問評価と設問ベース自動評価手法を用いて行った日中エンジン評価について説明する。テストセット評価（設問ベース評価）の自動化手法が従来の自動評価手法と比較してどの程度の信頼性を持っているか、また、設問回答がプログラムと人間でどの程度の一致が見られるかを調べることができるように、以下の要領で評価実験を行った。

① 人間による評価：

・評価者：2 名（日本語検定 1 級を持つ中国語ネイティブ）

・評価観点

－正確さ（Adequacy）：原文と訳文を見て 1～5 で評価

－流暢さ（Fluency） ：訳文のみを見て 1～5 で評価

－設問 ：訳文と設問を見て Yes/No 評価

② プログラムによる自動評価

－従来の自動評価尺度：BLEU、NIST、WER、PER

－自動化された設問評価

二人の評価者は、日本語のスキルが高く、かつ、言語学的な素養のある中国語ネィティブ（中国人）である。評価対象とした翻訳システムは、Web 上でアクセスすることができる 6 つの代表的な日中翻訳システムを選択した。6 つの翻訳システムを翻訳方式で分類すると、4 つが RBMT、2 つが SMT である。

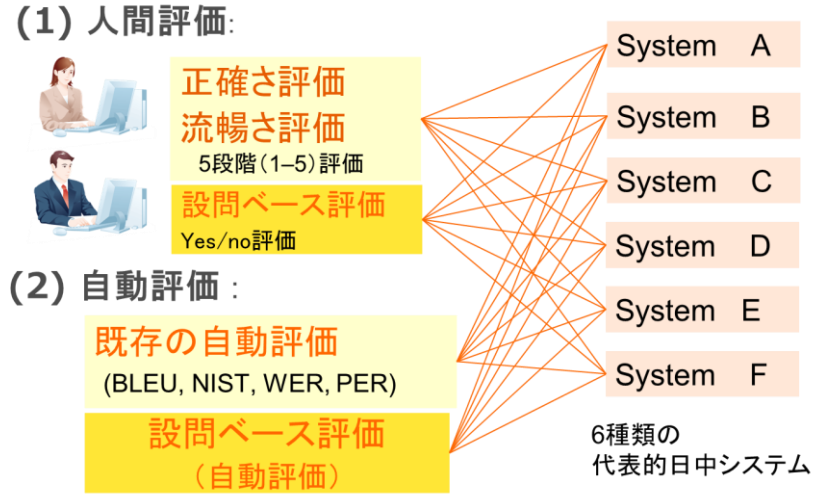


図 4 テストセットによる MT 評価実験

6.5.4.2 評価実験の結果と考察

テストセット評価が従来の自動評価尺度と比較してどの程度の信頼性を持っているかを検証するために、人間による主観評価（正確さ、流暢さ）との相関係数(Pearson product-moment correlation coefficient)を求める。正確さ、流暢さは、評価文ごとに評価者 1 と評価者 2 のつけたスコアを平均して、これを人間の評価値とした。同様に人間による設問評価についても、二人の評価者のつけたスコアの平均をとった。設問評価結果は、”Yes”回答を 1.0、”No”回答を 0.0 として数値化した。

６つの日中翻訳システムの従来自動評価（BLEU,NIST,WER,PER）によるスコア、設問評価（人間、設問）の平均値と、正確さ、流暢さの平均値の間の相関係数を表 2 に示す。全体的に相関係数が高い値になっているのは、通常のテストセットに比べて文が短いためと思われる。設問評価と正確さ（主観評価）との相関に関して、評価観点が異なるものの人間評価同士よりも人間評価対自動評価の方が高い値を示しているのは興味深い。

表２　各評価尺度と従来型主観評価との相関係数

	正確さ	流暢さ
BLEU	0.9453	0.9588
NIST	0.9783	0.9123
1-WER	0.9801	0.9267
1-PER	0.9707	0.8986
Questions（by Human）	0.9793	0.9334
Questions(Automatic)	0.9902	0.9208

表 2 から、いずれの評価尺度、評価方法においても、正確さ、および、流暢さとの間で高い相関を示している。相関の有意性を判定するために表２の相関係数を用いて t 検定を行い、いずれの組み合わせにおいても相関は有意水準 1%で有意であるという結果を得た。従来自動評価と設問評価の相関係数を比較すると、正確さにおいては設問評価の数値が従来自動評価尺度の数値を上回っている。このことから、設問による評価手法は人間による主観評価の代替手法として有効であり、かつ、従来自動評価尺度に劣らない信頼性を持っていると言える。
次に、設問評価において自動評価と人間評価の一致度を調べるために、それぞれの評価結果の間で Kappa 係数を求めた。その結果を表３に示す。有意水準 1%で検定を行ったところ、評価者１、評価者２のどちらにおいても人間と自動評価のつけるスコアは一致しているという結果になった。このことより、設問ベース自動評価を人間評価の代わりに用いても問題はないと言える。

表３　　自動評価と人手評価の間の Kappa 係数

自動評価:人手評価（評価者 1)	0.5494
自動評価: 人手評価(評価者 2)	0.5552
人手評価(評価者１)：人手評価(評価者 2)	0.6616

6.5.5　テストセット評価のメリットと課題

本節では、前節で紹介した実験結果を踏まえ、テストセット評価と他の翻訳評価法とを比較して、テストセット評価のメリットおよび課題について整理を行う。

6.5.5.1 テストセット評価のメリット

① 翻訳エンジンごとに文法項目別のフィードバックが得られ、評価結果を見てエラー分析ができる
② 機械翻訳開発ベンダにとって、自システムの長所や短所を具体的に知ることができ、次の開発に注力すべきポイントを決めることができる

これらの点はテストセット評価にユニークな特徴であり、最も注目すべきメリットであると言える。

③ テストセット評価は”Yes”または”No”の 2 値で回答するため、5 段階を基本とする人間評価に比べてコスト面、評価値の客観性の面からも有利である

従来型人間評価に比べて評価に必要な時間と手間が節減できると思われる。

④ テストセットの自動評価手法はテストセットの人間評価と強い相関があり、自動評価を人間評価の代わりに用いることができる

設問評価を自動評価化することによって、さらに低コストの評価が可能となる。

⑤ テストセット評価手法は従来型自動評価手法の代わりに使うことができるかもしれない
⑥ 設問を正規表現形式に書き換えてテストセット評価を自動化することが可能

テストセットの設問は文章の一部をピンポイントで質問するものだが、例文数を増やしていってチェックする文法項目を網羅することによって、BLEU など従来型自動評価の代替として機能する可能性がある。

6.5.5.2 テストセット評価の課題

① 言語対毎に対訳例文対、そして各例文に設問を設定しなければならない
② 設問の設定には言語的素養を持つ人が作業を行う必要がある

テストセットの開発には設問の設定が必要であり、従来自動評価に比べて準備のためのコストが大きい。テストセットを効率的に開発する方法の確立が課題である。

③ エンジンの実力とテストセットの例文のレベルが合わない場合、回答が Yes または No のどちらかに偏ってしまう
④ 標準的なテストセットだけでは、対象文書の分野や文種に特有の表現や語彙に関する情報

を得ることができない

原文の難易度別、分野別、文種別などの観点別にテストセットを複数用意しておき、翻訳エンジンの実力や対象文章の性質に合わせてテストセットを使い分けられるような環境が作れると理想的である。

⑤ テストセットに合わせてエンジンをチューニングすると同じテストセットを繰り返し使うことができない

テストセットをできるだけ多くの人に使ってもらいながら、テストセットへのチューニングが容易にできないような公開の仕方を考える必要がある。

6.5.6 テストセット評価の将来展望

テストセット評価のユニークな特徴である「エラー分析」のフィードバックをより詳細化するためには、課題③④で述べたように観点別にテストセットを複数用意することが望ましい。テストセット作成のトータル工数を抑えるために、テストセットを階層的に作成し組み合わせて利用するというアイデアがある。

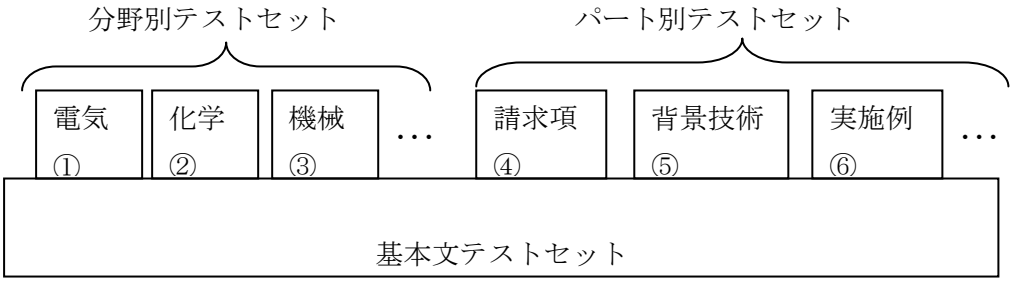


図５　階層的テストセットの例

図５は特許文書を評価対象とした場合の、階層的なテストセットの構成例である。すべての文章に共通の文法項目をチェックするための「基本文テストセット」をベースとし、この上に分野別、パート別というように、観点毎に特有な表現や語彙をチェックするための観点別テストセットを用意する。たとえば、電気分野の請求項を評価するときは、基本文テストセットと電気テストセット（①)、請求項テストセット（④）の３つを組み合わせて利用すればよい。こうすることで、観点のすべての組み合わせに対応して個別にテストセットを作る必要がなくなり、効率的な評価が可能になると考えられる。

テストセットの作成は、原言語と目標言語のスキルを持った専門家がひとつひとつ例文を集めて

設問を作るという作業が必要である。特許文書を対象とする場合には、例文収集作業を既存の特許データベースを利用して省力化できる可能性がある。国際特許出願には PCT（Patent Cooperation Treaty）と呼ばれる制度があり、PCT に則って複数の国に出願された特許には共通の番号が振られるようになっている。したがって、図６のように、その共通の番号をキーにして日本語特許と英語特許、日本語特許と中国語特許など言語が異なるが内容が一致している特許案件のペア（特許ファミリ）を各国の特許データベースから自動的に収集することができる。[5)]

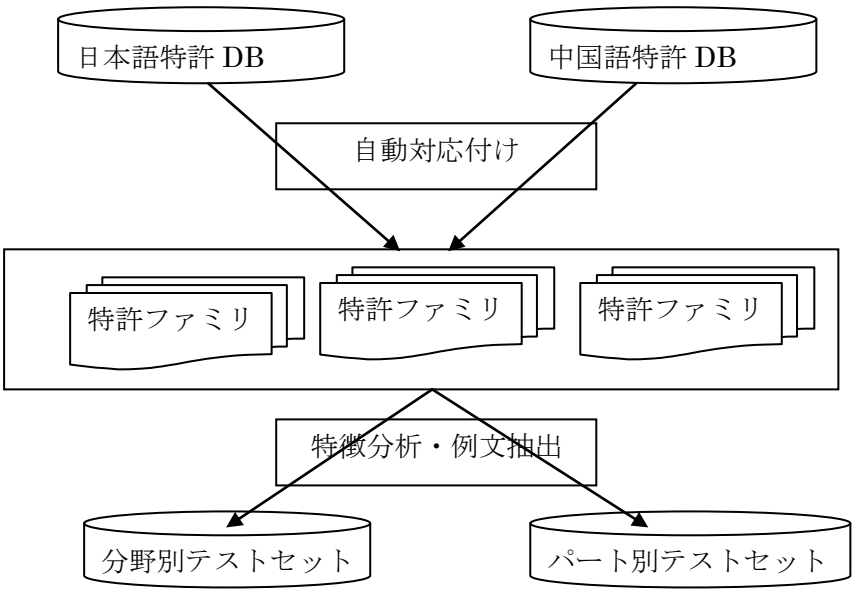


図６　効率的な特許文書用テストセットの作成

各特許には IPC と呼ばれる万国共通の特許分類番号が割り当てられているため、IPC を用いれば特許ファミリを分野別に分類することができる。また、特許明細書に含まれる項目ラベルを参照すれば、請求項、実施例などのパート別にテキストを取り出すことも容易である。したがって、特許ファミリ群から分野別、パート別の対訳テキストを自動抽出することは難しいことではない。さらに、分野別、パート別に自動抽出された対訳テキストを分析して、それぞれの特徴的な表現を含む例文を選択し、設問を付加することによって、分野別、パート別のテストセットが完成する。将来、分類された対訳テキストの特徴分析を自動化する研究が進めば、特許ファミリから全自動でテストセットを作成することが可能になるかもしれない。

6.5.7　おわりに

JEIDA－AAMT で開発してきた機械翻訳評価用テストセットを例にとりながら、テストセット評価について概観した。テストセット評価は、エラー分析ができるという他の評価方法にはない特徴を備えるほか、従来型人間評価との高い相関を持っており BLEU 等の自動評価の代用手法とし

ても期待される。しかし、テストセット作成のコストの高さや再利用性の問題などから、現在、実際に翻訳評価方法として用いられる例は稀である。
特許分野のように多観点でグルーピングされた対訳テキストが大量に抽出できる環境においては、近い将来、テストセット作成プロセスを自動化するための研究が進むことによって大幅なコストダウンが実現すると考えられる。評価用テストセットが容易に利用できる環境が整えば、翻訳の評価手法に新たな選択肢が加わることになり、より柔軟で効率的な翻訳評価ができるようになると確信している。

References

1) Papineni K., S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. *Proc. of ACL*, 311–318.
2) Isahara, H. 1995. JEIDA's Test-Sets for Quality Evaluation of MT Systems --Technical Evaluation from the Developer's Point of View--. *Proc. of MT Summit V*.
3) Uchimoto, K., K. Kotani, Y. Zhang and H. Isahara. 2007. Automatic Evaluation of Machine Translation Based on Rate of Accomplishment of Sub-goals. *Proc. of NAACL HLT*, 33-40.
4) Nagase, T., H. Tsukada, K. Kotani, N. Hatanaka and Y. Sakamoto. 2011. Automatic Error Analysis Based on Grammatical Questions . *Proc. of PACLIC*.
5) Higuchi, S. M. Fukui, A Fujii and T. Ishikawa. 2001. A system for multi-lingual patent retrieval. *Proc. Of MT Summit VIII*.

６．６　自動評価尺度 IMPACT の実用化に向けて

～処理時間短縮のための最適化手法～

北海学園大学　越前谷　博

6.6.1　はじめに

これまでに様々な自動評価尺度が提案されている。その中で深い言語知識に基づく尺度としては ULC[1]、MaxSim[2]、RTE[3]など、人手評価との間で高い相関が得られることが報告されている[4]。しかし、これらの自動評価尺度は高度な言語リソースに依存しているため、その利用は限定的である。一方、シンプルな単語マッチングに基づく尺度としては BLEU[5]、NIST[6]、PER[7]、WER[8]、TER[9]などがある。これらの自動評価尺度は言語リソースに依存しないため人手評価との相関において十分ではないが、利用が容易であり、また、処理時間も比較的短い。

これらのシンプルな単語マッチングに基づく自動評価尺度においては、更に、文単位の評価精度向上を目的にチャンクに基づく自動評価尺度である METEOR[10]、GTM[11]、ROUGE-W[12]、IMPACT[13]などが提案されている。チャンクとは 1 つ以上の一致単語からなる一致単語列を指す。IMPACT はこれらのチャンクに基づく自動評価尺度の中において評価精度が比較的高い[14]。IMPACT では、チャンクの長さと相対的な位置に基づき最適なチャンク列を決定している。しかし、その際に、DP（Dynamic Programming）に基づき全ての最長共通部分列（LCS）の経路を導き出し、その中から最適なチャンク列を決定するため、LCS の経路の候補が複数存在する場合、膨大な処理時間を要することとなり、実用化の点で大きな問題となる。そこで、本報告では、最適なチャンク列を効率よく決定するための最適化手法を提案する。

6.6.2　最適なチャンクの決定処理における問題点

チャンクの決定処理は通常、DP テーブルを用いて LCS を決定することで行う。DP テーブルを作成する際には以下の式を用いる。

$$D_{i,j} = \begin{cases} 0, & i = 0 \ or \ j = 0 \\ \max\left(D_{i-1,\,j}, D_{i,j-1}\right), & m_i \neq n_j \\ D_{i-1,\,j-1} + 1, & m_i = n_j \end{cases}$$

例えば、参照訳として “glass guide of the plastic mounting panel P”、MT 訳文として “a glass guide molded in panel member P made of the resin” が与えられた場合、DP テーブルは表 1 となる。表中の下線の数値は一致単語を示している。表 1 より、LCS の長さは 4 である。そして、LCS の長さが 4 である LCS 経路は LCS 経路 No.1: $D_{1,2}$ →$D_{2,3}$ →$D_{7,6}$ →$D_{8,8}$ と LCS 経路 No.2: $D_{1,2}$ →$D_{2,3}$ →$D_{3,10}$ →$D_{4,11}$ の 2 つである。それぞれの LCS 経路において、連続する一致単語をまとめて一つのチャンクとすることで参照訳と MT 訳文間の共通部分列は図 1 となる。図 1 の 2 つの LCS 経路の中から 1 つの LCS 経路を選択する場合、ROUGE-W や METEOR においては LCS 経路 No.2 が選択される。LCS 経路 No.2 のチャンク数は 2 であり、LCS 経路 No.1 のチャ

		j	1	2	3	4	5	6	7	8	9	10	11	12
		n_j	a	glass	guide	molded	in	panel	member	P	made	of	the	resin
i	m_i	0	0	0	0	0	0	0	0	0	0	0	0	0
1	glass	0	0	**1**	1	1	1	1	1	1	1	2	2	2
2	guide	0	0	1	**2**	2	2	2	2	2	2	2	2	2
3	of	0	0	1	2	2	2	2	2	2	2	**3**	3	3
4	the	0	0	1	2	2	2	2	2	2	2	3	**4**	4
5	plastic	0	0	1	2	2	2	2	2	2	2	3	4	4
6	mounting	0	0	1	2	2	2	2	2	2	2	3	4	4
7	panel	0	0	1	2	2	2	**3**	3	3	3	3	4	4
8	P	0	0	1	2	2	2	3	3	**4**	4	4	4	4

表 1　DP テーブルの例 1

LCS 経路 No.1: $D_{1,2}$ ->$D_{2,3}$ ->$D_{7,6}$ ->$D_{8,8}$
参照訳: [glass guide] of the plastic mounting [panel] [P]
MT訳文: a [glass guide] molded in [panel] member [P] made of the resin
LCS 経路 No.2: $D_{1,2}$ ->$D_{2,3}$ ->$D_{3,10}$ ->$D_{4,11}$
参照訳: [glass guide] [of the] plastic mounting panel P
MT訳文: a [glass guide] molded in panel member P made [of the] resin

図 1　DP テーブルに基づく共通部分列の例

ンク数 3 よりも小さい。即ち、LCS 経路 No.2 の方がチャンクの長さは大きいため、スコア値が大きくなる。また、チャンクを決定する際には、一致単語の連続性のみに着目するため膨大な処理時間を必要としない。

一方、IMPACT では、チャンクの長さだけでなく、相対的な位置も考慮した上で、チャンク列を決定する。具体的には以下の式を用い、score の値が大きい方の LCS 経路が選択される。

$$score = \sum_{c \in c\_num} \left(length(c)^{\beta} \times pos \right)$$

$$pos = \left(1.0 - \left| \frac{c_i}{m} - \frac{c_j}{n} \right| \right)$$

図 1 の参照訳と MT 訳文間においては、β の値が 1.2 の際には LCS 経路 No.1 の score は 3.4933(=$2^{1.2}$×(1.0-|1/8-2/12|)+$1^{1.2}$×(1.0-|7/8-6/12|)+$1^{1.2}$×(1.0-|8/8-8/12|))となり、LCS 経路 No2 の score は 3.4461(=$2^{1.2}$×(1.0-|1/8-2/12|)+$2^{1.2}$×(1.0-|3/8-10/12|))となる。したがって、score の値が大きい LCS 経路 No.1 が選択される。この場合、LCS 経路 No.1 のチャンクが人

間の直観により等しいと考えられる。ROUGE-W や METEOR は一致単語のチャンクの数のみに着目するため処理時間は短いが、最適なチャンク列の決定の点で不十分である。また、IMPACT は全ての LCS 経路を探索するため大幅な処理時間を要する。しかし、最適なチャンク列の決定においては有効である。したがって、ROUGE-W や METEOR のように処理時間が短く、かつ IMPACT のように最適なチャンク列を決定できる自動評価尺度が必要となる。このような観点より、本報告では、IMPACT を対象として、処理時間の短縮のための最適化手法を提案する。

6.6.3 最適なチャンク列決定のための最適化手法

提案手法では始めに DP テーブルを作成する。例えば、参照訳“array rules determine the limit to designing of the wiring routes”と MT 訳文“arrangement of restriction on the design rule , the wiring route is determined”の間において、表 2 の DP テーブルが生成される。

		j	1	2	3	4	5	6	7	8	9	10	11	12	13
		n_j	arrang ement	of	restrict ion	on	the	design	rule	,	the	wiring	route	Is	resin
i	m_i	0	0	0	0		0	0	0	0	0	0	0	0	0
1	array	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	rules	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	determine	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	the	0	0	0	0	0	1	1	1	1	1	1	1	1	1
5	limit	0	0	0	0	0	1	1	1	1	1	1	1	1	1
6	to	0	0	0	0	0	1	1	1	1	1	1	1	1	1
7	designing	0	0	0	0	0	1	1	1	1	1	1	1	1	1
8	of	0	0	1	1	1	1	1	1	1	1	1	1	1	1
9	the	0	0	1	1	1	2	2	2	2	2	2	2	2	2
10	wiring	0	0	1	1	1	2	2	2	2	3	3	3	3	3
11	routes	0	0	1	1	1	2	2	2	2	3	3	3	3	3

表 2　DP テーブルの例 2

表 2 に基づき LCS 経路の候補を決定した場合、3 つの LCS 経路の候補が存在し、それらは図 2 のようなチャンク列に相当する。従来の IMPACT では、3 つの LCS 経路の候補から LCS 経路 No.2 を選択する。

次いで、表 2 の DP テーブルの一致単語のみに着目した木構造を構築する。一致単語は $D_{8,2}$、$D_{9,5}$、$D_{10,10}$、$D_{4,5}$、$D_{9,9}$、$D_{4,9}$ の箇所であり、それぞれの箇所の $D_{i,j}$ の値は 1、2、3、1、2、1 である。そして、これらの値を降順に並べ、木構造のネストとする。その結果、図 3 に示すようなノードを持つ木構造が生成される。ノード間のリンクについては、連続する 2 つの Value の間のノードにおいて、上位のノードの i と j が下位のノードの i と j よりも共に大きい場合のみ張られる。したがって、図 3 のノード間においては、図 4 のようなリンクが張られる。図 3

LCS経路No.1: $D_{8,2}$ ->$D_{9,9}$ ->$D_{10,10}$

参照訳: array rules determine the limit to designing [of] [the wiring] routes

MT訳文: arrangement [of] restriction on the design rule , [the wiring] route is determined

LCS経路No.2: $D_{4,5}$ ->$D_{9,9}$ ->$D_{10,10}$

参照訳: array rules determine [the] limit to designing of [the wiring] routes

MT訳文: arrangement of restriction on [the] design rule , [the wiring] route is determined

LCS経路No.3: $D_{8,2}$ ->$D_{9,5}$ ->$D_{10,10}$

参照訳: array rules determine the limit to designing [of] [the] [wiring] routes

MT訳文: arrangement [of] restriction on [the] design rule , the [wiring] route is determined

図 2　DP テーブルに基づく LCS 経路の候補とそのチャンク列

Value	$D_{i,j}$
3	$D_{10,10}$
2	$D_{9,9}$　$D_{9,5}$
1	$D_{8,2}$　$D_{4,5}$　$D_{4,9}$

図 3　木構造におけるノードの決定

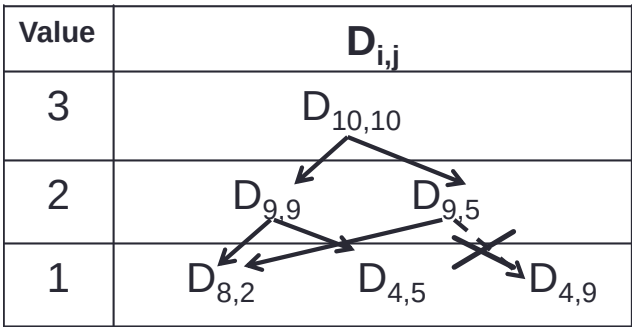


図 4　木構造におけるリンクの決定

においては、ノード $D_{9,5}$ と $D_{4,9}$ の間では j の値は下位のノードの値 9 が上位のノードの値 5 よりも大きいためリンクは張られない。その結果、いずれのノードともリンクが存在しないノード $D_{4,9}$ は削除される。表 2 のような DP テーブルをそのまま使用して LCS 経路の全候補を決定すると一致単語以外の単語間についても DP テーブルを走査しなければならいなため膨大な処理時間を要する。それに対して、DP テーブルを木構造に変換することにより、一致単語のみを対象として LCS 経路を決定するため処理時間を大幅に減少させることが可能となる。

更に、提案手法では最適な LCS 経路を効率的に決定するために木構造の近似を行う。

IMPACT では、チャンクの長さと相対的な位置関係に基づきチャンク列を決定している。即ち、木構造に置き換えると連続している 2 つのノード、または、i と j の値の差が小さなノードが優先されるということを意味する。そこで、2 つのノード間において下位のノードの i と j にそれぞれ 1 を加えた i と j を持つノードが上位ノードとして存在する場合にはそれらは連続しているため、必要なノードと位置付ける。更に、同じ Value に存在するノードの中で i と j の差が最小のノードのみを残す。この条件を取り入れることにより、図 4 の木構造は図 5 のように近似された木構造となる。

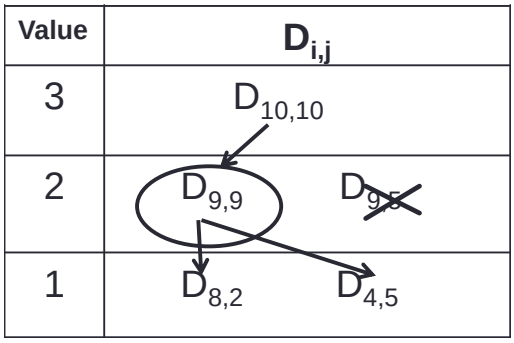


図 5　近似された木構造

図 4 の木構造に対して図 5 ではノード $D_{9,5}$ が削除される。Value が 2 である 2 つのノード $D_{9,9}$ と $D_{9,5}$ において、$D_{9,9}$ は Value が 3 の上位ノード $D_{10,10}$ との関係において連続している。したがって、ノード $D_{9,9}$ は 必要なノードとなり Value の 2 に残る。ノード $D_{9,5}$ は、上位、下位共に連続するノードは存在しない。また、ノード $D_{9,9}$ と $D_{9,5}$ の i と j の相対的な位置のズレはそれぞれ 0.1259(=|9/11-9/13|)と 0.4336(=|9/11-5/13|)になり、ノード $D_{9,9}$ の方が相対的な位置のズレは小さい。したがって、ノード $D_{9,5}$ は木構造から削除される。その結果、図 2 の 3 つの LCS 経路の候補の中からノード $D_{9,5}$ を含む LCS 経路 No.3 が削除される。そして、残された LCS 経路 No.1 と LCS 経路 No.2 に対して、IMPACT と同様に式を用いて、最適なチャンク列として LCS 経路 No.2 を一意に決定する。このように必要なノードのみを残すことにより、不要な LCS 経路が削除され、処理時間の短縮をもたらすと考えられる。

6.6.4　スコア値の算出

一意に決定された LCS 経路に基づき自動評価としてのスコア値を算出する。その際には以下の計算式を用いる。

$$R = \left(\frac{\sum_{i=0}^{RN} \left(\alpha^i \sum_{c \in c\_num} length(c)^\beta \right)}{n^\beta} \right)^{\frac{1}{\beta}}$$

$$P = \left(\frac{\sum_{i=0}^{RN} \left(\alpha^i \sum_{c \in c\_num} length(c)^\beta \right)}{m^\beta} \right)^{\frac{1}{\beta}}$$

6.6.3 の処理より図 2 の LCS 経路の候補の中の LCS 経路 No.2 が一意に決定される。この

LCS 経路 No.2 のチャンクは上式を用いることにより図 6 に示すように数値化される。ここで、パラメータ β は 2.0 を用いている。チャンク列 No.1 においては、“the”と“the wiring”の 2 つのチャンクに基づき 5 が得られる。更に、チャンク列を再帰的に決定することでチャンク列 No.2 が得られ、このチャンク列に対しては 1 が付与される。

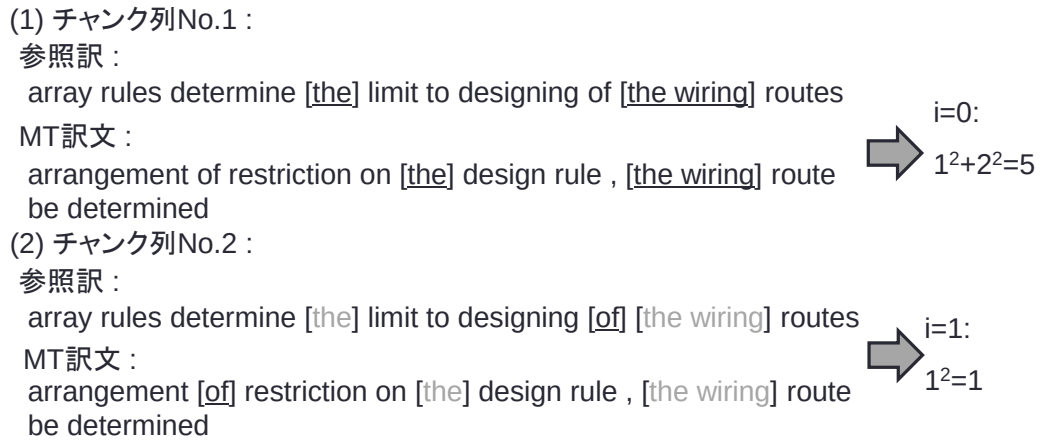


図 6　チャンク列の数値化の例

次いで、カウンタ i をスコア計算に反映させる。ここで、パラメータ α には 0.5 を用いている。チャンク列 No.1 においては 5.0(=$0.5^0\times5$)、チャンク列 No.2 においては 0.5(=$0.5^1\times1$)となり、この 2 つの数値の和である 5.5 が上式 R と P の分子の値となる。R と P の分母においては R が参照訳の単語数を表しているため $11^{2.0}$ となり、P は MT 訳文の単語数を表しているため、$13^{2.0}$ となる。したがって、R と P の値はそれぞれ 0.2132、0.1804 となる。そして、以下の式を用いて最終的なスコア値を算出する。その結果、γ の値は 0.8462(=0.1804/0.2132)となり、score の値は 0.1928 となる。

$$score = \frac{(1+\gamma^2)PR}{\gamma^2 P + R}$$

$$\gamma = \frac{P}{R}$$

6.6.5　性能評価実験

提案手法の有効性を確認するために性能評価実験を行った。実験データとしては NTCIR-7[15]の特許翻訳データを用いた。14 の機械翻訳システムが日本語を英語に翻訳した訳文 100 文ずつの計 1,400 文を MT 訳文に用いた。また、訳文 100 に対して 4 つの参照訳を用いた。これらのデータを用いて、始めに処理時間についての評価を行った。提案手法と IMPACT の処理時間における比較実験を表 3 に示す。表 3 より全データ 1,400 の MT 訳文を用いた場合には 5 秒の短縮が確認された。それほど大きな処理時間の短縮ではないが、LCS 経路の候補数が少ない場合、IMPACT でもそれほど多くの処理時間を必要としない。提案手法が有効に働くのは、

	1,400文	**LCS経路の候補数が300以上であった8文**
提案手法	144 sec	5 sec
IMPACT	149 sec	7 sec

表 3　処理時間における実験結果

LCS 経路の候補数が多い場合である。今回のデータでは、そのほとんどが LCS 経路の候補数が 100 以下であった。そこで、LCS 経路の候補数が 300 以上であった 8 文のみを選択して、処理時間を比較した。その結果、8 文だけにもかかわらず、処理時間が 7 秒から 5 秒と 2 秒短縮された。この結果は、処理時間の短縮を目的とした提案手法の有効性を示すものである。

更に、提案手法の自動評価尺度の精度に対する影響について調査を行った。調査方法は、IMPACT と提案手法において人手評価との相関を求めた。人手評価は Adequacy と Fluency の観点より 3 名のバイリンガルが 5 段階評価を行った結果を用いた。各 MT 訳文の人手評価には 3 名の評価値のメジアン値を用いた。相関は Adequacy と Fluency の両方に対して、ピアソンの相関係数とスピアマンの順位相関係数を求めた。自動評価尺度としては提案手法、IMPACT、そして、ROUGE-W を用いた。表 4 には Adequacy におけるピアソンの相関係数を示す。表 5 には Fluency におけるピアソンの相関係数を示す。

Metric	No.1	No.2	No.3	No.4	No.5	No.6	No.7	No.8
提案手法	0.7625	0.5307	0.4704	0.5566	0.5518	0.6295	0.6516	0.7375
IMPACT	0.7625	0.5307	0.4704	0.5566	0.5518	0.6295	0.6516	0.7374
ROUGE-W	0.7648	0.5044	0.4615	0.5765	0.5482	0.6257	0.6415	0.7336
Metric	No.9	No.10	No.11	No.12	No.13	No.14	Avg.	System
提案手法	0.7113	0.5813	0.7095	0.6251	0.7677	0.5321	0.6298	0.9266
IMPACT	0.7113	0.5813	0.7097	0.6251	0.7685	0.5321	0.6299	0.9264
ROUGE-W	0.7072	0.5938	0.7131	0.6099	0.7643	0.5402	0.6275	0.9400

表 4　Adequacy におけるピアソンの相関係数

Metric	No.1	No.2	No.3	No.4	No.5	No.6	No.7	No.8
提案手法	0.5543	0.3765	0.3705	0.5548	0.4737	0.5786	0.5168	0.6968
IMPACT	0.5543	0.3765	0.3705	0.5548	0.4737	0.5786	0.5168	0.6968
ROUGE-W	0.5566	0.3501	0.3504	0.5715	0.4693	0.5791	0.5006	0.6941
Metric	No.9	No.10	No.11	No.12	No.13	No.14	Avg.	System
提案手法	0.5564	0.5514	0.6333	0.3727	0.6081	0.4012	0.5175	0.9382
IMPACT	0.5564	0.5514	0.6333	0.3728	0.6081	0.4012	0.5175	0.9381
ROUGE-W	0.5480	0.5520	0.6410	0.3568	0.6003	0.4078	0.5127	0.9426

表 5　Fluency におけるピアソンの相関係数

また、表 6 には Adequacy におけるスピアマンの順位相関係数を示す。表 7 には Fluency におけるスピアマンの順位相関係数を示す。表中の Avg.は MT システム No.1 から MT システム No.14 までの文単位での相関係数の平均を表している。System は、システムレベルの相関係数である。

Metric	No.1	No.2	No.3	No.4	No.5	No.6	No.7	No.8
提案手法	0.7468	0.4618	0.4923	0.5666	0.4877	0.6280	0.6196	0.6464
IMPACT	0.7468	0.4618	0.4923	0.5666	0.4877	0.6280	0.6196	0.6460
ROUGE-W	0.7379	0.4494	0.4943	0.5786	0.4785	0.6166	0.5902	0.6375
Metric	No.9	No.10	No.11	No.12	No.13	No.14	Avg.	System
提案手法	0.6859	0.5478	0.7171	0.5960	0.7433	0.5836	0.6088	0.9912
IMPACT	0.6859	0.5478	0.7171	0.5971	0.7442	0.5836	0.6089	0.9912
ROUGE-W	0.6734	0.5653	0.7195	0.5737	0.7448	0.5682	0.6020	0.9912

表 6　Adequacy におけるスピアマンの順位相関係数

Metric	No.1	No.2	No.3	No.4	No.5	No.6	No.7	No.8
提案手法	0.5520	0.3518	0.3643	0.5488	0.4119	0.5995	0.4691	0.6321
IMPACT	0.5520	0.3518	0.3643	0.5488	0.4119	0.5995	0.4691	0.6325
ROUGE-W	0.5566	0.3501	0.3504	0.5715	0.4693	0.5791	0.5006	0.6941
Metric	No.9	No.10	No.11	No.12	No.13	No.14	Avg.	System
提案手法	0.5452	0.4661	0.6548	0.3590	0.6371	0.4437	0.5025	0.9253
IMPACT	0.5452	0.4661	0.6544	0.3586	0.6359	0.4437	0.5024	0.9253
ROUGE-W	0.5480	0.5520	0.6410	0.3568	0.6003	0.4078	0.5127	0.9426

表 7　Fluency におけるスピアマンの順位相関係数

表 4 から表 7 より、Avg.と System において提案手法と IMPACT との差は 0.0001 から 0.0002 の間と非常に小さなものであった。したがって、提案手法による自動評価としての影響は僅かであり、精度を落とすことなく処理時間の短縮を実現できていることが明らかとなった。

6.6.6　まとめ

本報告では、チャンクに基づく自動評価尺度において最適なチャンクを決定すると共に処理時間の短縮が可能な最適化手法を提案した。そして、提案手法を自動評価尺度 IMPACT に適用し、その有効性を確認した。提案手法は翻訳自動評価としての精度を低下させることなく、処理時間の短縮を実現させるものであった。今後は、自動評価の精度向上のための改良を行う予定である。

謝辞

この研究は国立情報学研究所との共同研究に関連して行われた。

参考文献

[1] Jesús Giménez and Lluís Márquez(2007) “Heterogeneous Automatic MT Evaluation Through Non-Parametric Metric Combinations,” Proceedings of IJCNLP, pp. 319-326.

[2] Yee Seng Chan and Hwee Tou Ng(2008) “MAXSIM: An Automatic Metric for Machine Translation Evaluation Based on Maximum Similarity,” Proceedings of the Metrics-MATR Workshop of AMTA-2008, pp. 319-326.

[3] Sebastian Padó, Michel Galley, Dan Jurafsky, Christopher D. Manning(2009) “Textual Entailment Features for Machine Translation Evaluation,” Proceedings of the 4th Workshop on Statistical Machine Translation.

[4] Chris Callison-Burch, Philipp Koehn, Christof Monz and Josh Schroeder(2009) “Findings of the 2009 Workshop on Statistical Machine Translation,” Proceedings of the Fourth Workshop on Statistical Machine Translation, pp.1-28.

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu(2002) “BLEU: a Method for Automatic Evaluation of Machine Translation,” Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311-318.

[6] NIST(2002) “Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics”

[7] Keh-Yih Su Ming-Wen Wu and Jing-Shin Chang(1992) “A New Quantitative Quality Measure for Machine Translation Systems,” Proceedings of the 14th International Conference on Computational Linguistics, pp.433-439.

[8] Geroge Leusch, Nicola Ueffing and Hermann Ney(2003) “A Novel String-to-String Distance Measure With Applications to Machine Translation Evaluation,” Proc. of MT Summit IX, pp.240-247.

[9] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla and John Makhoul(2006) “A Study of Translation Edit Rate with Targeted Human Annotation,” Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA), pp. 223-231.

[10] Alon Lavie and Abhaya Agarwal(2007) “Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments,” Proceedings of the Second Workshop on Statistical Machine Translation, pp. 228–231.

[11] Joseph P. Turian, Luke Shen and I. Dan Melamed(2003) “Evaluation of Machine Translation and its Evaluation,” Proc. of MT Summit IX, pp.386-393.

[12] Chin-Yew Lin and Franz Josef Och(2004) “Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics,” Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311-318.

[13] Hiroshi Echizen-ya and Kenji Araki(2007) “Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum,” Proceedings of the Eleventh Machine Translation Summit, pp.151-158.
[14] Hiroshi Echizen-ya, Terumasa Ehara, Sayori Shimohata, Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Noriko Kando(2009) “Meta-Evaluation of Automatic Evaluation Methods for Machine Translation using Patent Translation Data in NTCIR-7,” Proceedings of the 3rd Workshop on Patent Translation, pp.9-16.
[15] Fujii, Atsushi., Utiyama, Masao., Yamamoto, Mikio. and Utsuro, Takehito(2008) “ Overview of the Patent Translation Task at the NTCIR-7 Workshop,” Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, pp.389-400.

7．　第２回特許情報シンポジウム報告

山形大学　横山 晶一

（本稿は、AAMT ジャーナル 53 号に掲載のものと同じである。転載を許可された AAMT に感謝する）

7.1　はじめに

AAMT/Japio 特許翻訳研究会（委員長：辻井潤一東大名誉教授）では、機械翻訳国際会議(MT Summit)のない年に、国内でさまざまなイベントを行っている。

特許情報シンポジウムは、2010 年（平成 22 年）に第１回が東大・本郷キャンパスで行われ、今回 2012 年（平成 24 年）に第２回が行われた。ここでは、その内容等について簡単に報告する。

7.2　シンポジウムの開催趣旨

論文募集には、次のような開催趣旨が書かれている。

「特許情報は情報処理技術の応用分野の一つとして近年世界的に関心が高まっている。本シンポジウムは、特許情報処理技術の研究開発を促進することを目的として、2010 年に第１回が開催された。研究者、実務家、政府関係者が集まって、構想、方法論、将来展望、実務経験、政策などを議論する場として企画され、さまざまな発表が行われた。今回第 2 回のシンポジウムも、第１回の成果と趣旨を踏まえて、招待講演（3 件）、検討会報告（1 件）、一般講演（5 ないし 10 件）で構成される。」

この研究会がふだん取り扱っている特許翻訳の問題よりはやや幅広く、特許情報全般について取り扱う意志のもとに企画されている。

上記のように、このシンポジウムは、招待講演や報告とともに、一般に公募した論文発表も行われる。一般公募の論文の領域は、上記趣旨では次のようになっている。

論文の分野

下記の分野を含む特許情報処理の諸分野:

- Analysis and classification for patent documents,
- Machine translation and translation aids for patent documents,
- Contrastive studies for multilingual patent documents,
- Language resources for patent documents,
- Dictionaries and terminology databases for patent documents,
- Parallel, comparable or monolingual corpora for patent documents,
- Information extraction and information mining from patent documents,
- Patent map development,
- Evaluation techniques for patent translation,
- Patent information retrieval.

今回は締め切りまでに４編の論文投稿があった。

7.3 開催概要

開催日時：平成 24 年 11 月 30 日（金）

13：00～18：00

場所：京都大学東京オフィス（東京都港区）

参加者：事前登録車 98 名、当日登録者 9 名、欠席者 20 名、当日参加者 87 名

会場の関係で、100 名限定で参加を受け付け、ほぼ満席となる盛況であった。また、昨今の国際情勢から、海外からの参加がやや懸念されたが、特に問題なく参加していただいた。

次節で、実際の内容について述べる。

7.4 開催内容

今回は、海外からの参加者のことを考慮して、一般講演以外は英語で行われ、一般講演は、英語、日本語どちらでもよいことにしたが、４件の発表すべて日本語で行われた。

最初に、主催者側として、守屋敏道氏（Japio 専務理事、特許情報研究所所長）と辻井委員長から開会のあいさつがあった。

招待講演は３件行われた。

最初は、China Patent Information Center (CPIC)の王丹氏による“Making Effective Use of Machine Translation for Patent Documents: Practice of CPIC”という講演であった。ここで氏は、CPIC で特許文書を扱う際の MT 使用ロードマップを示すとともに、人間の翻訳と MT との使い分けなどにも言及した。ロードマップでは、最近４～５年間で、中英、英中、日中、中日などの MT 使用や開発・研究が行われていることを述べた。

２番目の招待講演は、Paul Schwander 氏 (Director External Products and Services European Patent Office, EPO)による“Machine Translation at the EPO Removing language barriers from patent documentation”である。Schwender 氏は、EU 内での 23 言語に関する取り組みと、最近 EU 内でも需要が高まっているアジア系言語による特許への調査に MT がどのように寄与しているかについての現状を述べた。

３番目の稲葉崇氏（特許庁総務部普及支援課情報企画室調査第二係長）による「特許庁における機械翻訳に関する取組」（英文タイトルは JPO's Approach for Machine Translation - To establish productive utilization method and create appropriate policies）である。この講演では、特許庁における MT 利用の現状、特に、MT の精度向上や、MT を活用した外国文献へのアクセス向上の取組、MT の品質評価について述べた。

次に、9 月 7 日に開かれた特許文書の機械翻訳結果評価方法検討会（東大本郷キャンパス、このジャーナルで概要を既報）のまとめを、このシンポジウムのまとめ役となった当研究会の江原暉将副委員長（山梨英和大教授）、越前谷博委員（北海学園大准教授）が行った。MT 評価に用いられる BLEU, NIST, RIBES, IMPACT といった自動評価や、人間による評価について議論をしたシンポジウムを総括し、今後の指針を示した。

一般発表は次の４件である。

「依存関係を用いた特許分野のための日英中対訳フレーズの切り出しアルゴリズム」池田秀人氏ら立命館大グループによる発表。依存関係階層木の中間節点を発展させた対訳フレーズを用

いて MT を行おうという試みである。

「特許明細書の翻訳者から基本的な誤訳の実例を示して対策を提案」吉川潔氏（フリー翻訳者）。氏が長年にわたる特許翻訳の経験から、MT における種々の誤謬について持論を述べたものである。

「特許翻訳の品質を向上させるための形態素解析結果を利用した文書比較・日本語精査ツール-歌詠と鶯-の試作」楠本浩二氏（（株）クレステック）らのグループによる発表。特許翻訳品質向上のためのパーソナルツールとして､類似ファイルの差分などを提示するシステムについて述べた。

「技術調査のための特許情報抽出」太田貴久氏（豊橋技科大）らのグループによる発表。技術調査に必要な技術の総称、発明の作用などを高精度に取り出すことで、ユーザに有用な情報抽出を行うシステムについて述べた。

最後に簡単にまとめが行われた。

7.5 終わりに

以上が第２回特許情報シンポジウム報告である。主催者側で活躍された委員の皆さん、講演者に対する種々の交渉に当たられた Japio の方々、会場をお貸し下さった京都大学、事務を担当されたナビックスの担当者に感謝します。

今後もこのような催しを積極的に実施していきたいと考えております。

なお、シンポジウムの内容については、当研究会の HP (http://aantjapio.com/)上で公開の予定です。

Memo

平成２４年度AAMT/Japio特許翻訳研究会報告書
（機械翻訳及び辞書構築に関する研究及びシンポジウム・拡大評価部会報告）

発 行 日　平成25年3月

発　　行　一般財団法人 日本特許情報機構（Japio）
〒135-0016 東京都江東区東陽4丁目1番7号
佐藤ダイヤビルディング
TEL:(03) 3615-5511　FAX:(03) 3615-5521

編　　集　アジア太平洋機械翻訳協会（AAMT）

印　　刷　株式会社 ナビックス