翻訳の自動評価指標を用いた言語横断検索の性能予測

岸田和明（慶應義塾大学文学部）
kishida@slis.keio.ac.jp

1.はじめに

言語横断検索の場合，通常，検索質問を翻訳し，検索対象文書の言語に合わせてから照合作業を行うが，当然，この翻訳の品質がその検索性能に大きな影響を与えることになる。実際，Kishida, Kuriyama, Kando & Eguchi[1)] は，NTCIR テストコレクションでの言語横断検索のデータに対して回帰分析を試み，この点を実証的に確認している。しかし，この研究における翻訳の質の測定方法は非常に特殊な規則に基づいており，拡張性に乏しい。

そこで本研究では，この点を解決するために，基本的には Kishida ら [1)] の枠組みに依拠しつつ，機械翻訳の分野で最近活発に研究されている翻訳結果の自動評価指標を導入した分析を試みる。具体的には，英語からイタリア語への横断検索を事例とし，CLEF[☆1] のテストコレクションを使った検索実験を通じて，分析を進める。

2.検索性能を予測するための回帰モデル

検索質問の翻訳に基づく言語横断検索の性能に影響を及ぼす要因としては，

① 検索質問の翻訳の品質
② 検索質問自体の難易度

の 2 つが考えられる。それぞれの要因を変数 t と v とで表わせば，被説明変数である検索性能を y として，回帰モデルは，

$$y = a + bv + ct, \qquad (1)$$

となる。ここで，a，b，c はパラメータである。なお，この式は，Kishida ら [1)] が使用したものと同一である。

被説明変数の実際の測定には，平均精度（average precision）を利用する。具体的には，後述の検索システムを使って，言語横断検索を実行し，検索課題ごとに算出される平均精度を被説明変数として用いる。

一方，検索質問自体の難易度を測定するには，単言語検索の性能を利用することが考えられる。通常，言語横断検索用のテストコレクションにおける検索課題は，人手によって各言語に翻訳されている。そこで，今回の場合，イタリア語の検索課題を用いた単言語検索の平均精度を変数 v として用いることとする。

3.翻訳の品質の自動評価

3.1 BLEU

BLEU (Papineni ら [2)])は，機械翻訳の自動的な評価のために提案された指標である。BLEU の提案以降，この種の研究が進み，類似した指標もいくつか提案されている。しかし，それらは BLEU の性能を抜本的に改善するものではないことから，今回は，とりあえず BLEU のみを取り上げる。

BLEU は，機械翻訳文と正解文（人手による翻訳で，複数可）とを比較し，基本的には，機械翻訳文中の n グラム（単語ベース）が正解文中にどの程度出現するかを測定する。具体的には，

c_j^n :翻訳文中の j 番目の n グラムの出現数
$(j = 1,..., M_n ; n = 1,..., H)$

s_{jk}^n :翻訳文中の j 番目の n グラムが，k 番目の正解文中に出現する回数
$(k = 1,..., L)$

と定義して，$n = 1,..., H$ について，まず，

☆1 http://www.clef-campaign.org/

$$p_n = \frac{\sum_{j=1}^{M_n} \min(c_j^n, \max_{k=1,\ldots,L} s_{jk}^n)}{\sum_{j=1}^{M_n} c_j^n},$$

を計算する。最終的に BLEU は,

$$BLEU(H) = BP \times \exp\left(\sum_{n=1}^{H} \frac{1}{H} \log p_n\right)$$

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

として算出される [2]。ここで c は機械翻訳文中の語の総数，r は“effective reference corpus length”[2]である。

3.2 WAMU

言語横断検索の場合，一般的な非専門用語よりも，専門的な用語がうまく翻訳されている場合のほうが，検索性能が高くなると予想される。そこで, 本研究では, BLEU の枠組みを少し変えた, idf による重み付け平均としての評価指標 WAMU (weighted average for matched unigrams)を提案する。すなわち,

$$WAMU = \sum_{j=1}^{M_1} w_j \frac{\min(c_j^1, \max_{k=1,\ldots,L} s_{jk}^1)}{c_j^1} \Bigg/ \sum_{j=1}^{M_1} w_j$$

である。ここで，重みは，idf として

$$w_j = \log \frac{N}{\tilde{n}_j}$$

で定義する。なお，$\tilde{n}_j$ は翻訳文中の j 番目のユニグラム（n=1）を含んだ文書数，N はデータベース中の全文書数である。

なお，上の式は n=1 に限定している。もちろんバイグラムでも idf は算出可能であるが，処理を必要以上に複雑にすることを避けるため，今回は WAMU の場合，常に n=1 とおく。

3.3 正解文の数

本研究では，テストコレクションとして提供されているイタリア語での検索課題を正解文として利用するため，常に正解文は 1 つである（上の式では常に k=1）。

4.検索実験と回帰分析

4.1 検索実験

検索実験に使用したのは，CLEF2003[3] におけるイタリア語文書集合（新聞記事の全文 157,558 件）と，英語およびイタリア語の検索課題集合である。検索課題は基本的には 60 件用意されているが, このうち, イタリア語文書集合中に適合文書をもつ 51 件の課題を今回は利用した（したがって, 以下の回帰分析等においては標本の大きさは常に 51 である）。

検索モデルとしては Okapi の BM25 を使い，英語からイタリア語への翻訳には，市販の機械翻訳（MT）ソフトウェア☆2 と対訳辞書☆3 による置換との 2 種類の方法を用いた。テキスト処理については，空白を目印に語分割し，標準的な方法で不要語の除去と語幹抽出を行った☆4。質問拡張などの特別な工夫は一切加えていない。

表 1　検索性能（MAP）

	単言語	言語横断	
		MT	辞書
MAP の値	0.386	0.326	0.296
単言語検索との比較：t 値	-	2.23*	2.90**

注　*：p<0.05，**：p<0.01

表 2 BLEU と WAMU との相関および基本統計量（MT の場合）

	BLEU(1)	BLEU(2)	WAMU
BLEU(1)	1.000	-	-
BLEU(2)	0.795	1.000	-
WAMU	0.808	0.594	1.000
平均	0.569	0.392	0.603
標準偏差	0.160	0.178	0.163
最大値	0.895	0.866	1.000
最小値	0.239	0.019	0.331

☆2 http://www.crosslanguage.co.jp/
☆3 http://www.freelang.net/
☆4 http://snowball.tartarus.org/

検索実験の結果を表 1 に示す。表 1 は平均精度の平均（MAP）の値であり，単言語検索と言語横断検索との間には検索性能に関して有意な差が認められる。

4.2 回帰分析

MT による翻訳結果に対する BLEU と WAMU の相関係数および基本統計を表 2 に示す。今回，BLEU については，ユニグラム（BLEU(1), *H*=1）とユニグラム・バイグラム（BLEU(2), *H*=2）とを計算した。その結果，ユニグラムの BLEU(1)と WAMU には高い相関がある一方，BLEU(2)と WAMU と間の相関係数はそれほど高くはなかった。

次に，MT に関する回帰分析の結果を表 3 に示す。予想通り，わずかな差異ではあるが，WAMU を用いた場合の回帰モデルの決定係数が最も高い。この回帰モデルは言語横断検索の性能の変動の約 67%を説明している。一方，バイグラム BLEU(2)の場合には，モデルの決定係数自体は低くはないが，BLEU(2)の回帰係数は-0.0417 であり，予測力をもたないことがわかる（P 値もかなり大きい）。

表 3　回帰分析の結果（MT）

(a)BLEU(1)　決定係数：0.6667

	回帰係数	標準誤差	P 値
切片	-0.0863	0.1032	0.4072
難易度	0.8317	0.0851	0.0000
BLEU(1)	0.1610	0.1643	0.3321

(b)BLEU(2)　決定係数：0.6605

	回帰係数	標準誤差	P 値
切片	0.0234	0.0747	0.7556
難易度	0.8369	0.0862	0.0000
BLEU(2)	-0.0417	0.1490	0.7809

(c)WAMU　決定係数：0.6671

	回帰係数	標準誤差	P 値
切片	-0.0906	0.1044	0.3895
難易度	0.8259	0.0851	0.0000
WAMU	0.1625	0.1606	0.3166

対訳辞書を使った場合の回帰分析の結果を表 4 に示す（BLEU(2)については省略）。この場合には，BLEU(1)を使ったモデルの決定係数が約 0.60 であるのに対して，WAMU によるモデルでは約 0.64 であり，WAMU の優位性がやや顕著に現れている。また，BLEU(1)と WAMU の P 値はともに低く，MT の場合よりも，予測因子としてこれらがより有効であることが示されている（WAMU の場合，回帰係数がゼロという帰無仮説が有意水準 1%で棄却されている）。

表 4　回帰分析の結果（対訳辞書）

(a)BLEU(1)　決定係数：0.6014

	回帰係数	標準誤差	P 値
切片	-0.1397	0.0847	0.1054
難易度	0.7621	0.0969	0.0000
BLEU(1)	0.3910	0.2119	0.0712

(b)WAMU　決定係数：0.6376

	回帰係数	標準誤差	P 値
切片	-0.1866	0.0753	0.0168
難易度	0.6839	0.0984	0.0000
WAMU	0.4256	0.1457	0.0053

なお，対訳辞書を用いた場合の WAMU でのモデルのみに関して，(1)式の両辺を対数変換した場合の結果を示しておく（表 5）。この場合，決定係数は低下するが，WAMU の予測因子としての有効性は高まっている。

表 5　対数回帰モデルの分析（対訳辞書）

	回帰係数	標準誤差	P 値
切片	0.1245	0.3505	0.7239
難易度	0.5914	0.1323	0.0000
WAMU	1.8547	0.3203	0.0000

注：決定係数は 0.5593

以上の分析の結果から，MT・対訳辞書の両方で，回帰モデル(1)式が言語横断検索の変動の約 60%以上を説明すること，また，翻訳の品質を測定する指標に関しては，WAMU のほうが BLEU よりも，予測因子

としての有効性がわずかに高いことが明らかとなった。

5.予測因子の実システムへの応用

翻訳の品質予測を言語横断検索システムに応用できれば便利である。例えば，利用者からの検索質問に対する翻訳の品質が低いと予測された場合に，検索質問の再入力を促す，あるいは，何らかのヘルプ機能で支援するなどのシステム上の工夫が考えられる。

残念ながら、表 3〜5 に示された回帰モデルをこの目的に応用することはできない。なぜなら，現実の検索状況では，「正解文」は与えられないからである。

正解文がないという状況で WAMU を算出する方法として，「再翻訳」の利用が考えられる。すなわち，本研究の状況ならば，英語から翻訳されたイタリア語の検索質問を，再度，英語に訳し戻し，その翻訳結果と元の英文との間で WAMU を計算すればよい。多くの MT システムでは，2 つの言語の双方向的な翻訳を可能としており，この点で，この方法は現実的である。

表 6　再翻訳の結果に対する回帰分析

	回帰係数	標準誤差	P 値
切片	-0.1676	0.0948	0.0833
難易度	0.8651	0.0843	0.0000
WAMU	0.2336	0.1151	0.0480

注：MT を利用，決定係数は 0.6868

もちろん，これによって計算される WAMU の値は 1 つの近似値にすぎず，検索対象文書の言語への翻訳結果の品質の程度と，再翻訳の結果のそれとが常に一致するという保証はない。しかし，例えば，元の検索質問に多義的な用語が含まれていた場合，その翻訳結果に曖昧性が生じ，結果的に，その再翻訳と元の検索質問とに差が生じるといった状況は考えうる。このような場合には，再翻訳の品質が最初の翻訳の品質を近似する可能性がある。

元の検索質問（英語）を正解文として，英語への再翻訳結果に対して計算した WAMU の値による回帰分析の結果を表 6 に示す。決定係数は約 0.69 で表 3(c)よりも高い。また，WAMU の P 値は 0.0480 で，帰無仮説は有意水準 5%で棄却される。この結果は，再翻訳に対する評価の実用可能性を示している。なお，表 3(c)で算出した WAMU と表 6 の WAMU との相関係数は約 0.289 であった。

6.おわりに

本研究では，翻訳の自動評価指標を用いて，言語横断検索の性能を回帰モデルで予測することを試みた。その結果，本研究で提案した WAMU を使い，単言語検索の性能を加えると，そのモデルはかなり高い予測力をもつことがわかった。

さらに，現実のシステムにこのモデルを適用する方法として，元の言語への再翻訳を利用できるかどうかの検討も試みた。

参照文献

1) Kishida, K.; Kuriyama, K.; Kando, N.; Eguchi, K. Prediction of performance on cross-lingual information retrieval by regression models. Proceedings of NTCIR-4. Tokyo, National Institute of Informatics, 2004.
2) Papineni, K.; Roukos, S.; Ward, T.; Zhu, W-J. BLEU: a method for automatic evaluation of machine translation. IBM Research Report RC22176 (W0109-022), 2001.
3) Braschler, M.; Peters, C. CLEF2003 methodology and metrics. C. Peters et al., eds., Comparative Evaluation of Multilingual Information Access Systems. Springer LCNS 3237. Berlin, Springer, 2004, p.7-20.