機械翻訳評価のための項目反応理論に基づく一対比較結果の統合

IRT-based Aggregation Model of Pairwise Comparison for Evaluating Machine Translations

京都大学大学院情報学研究科
大谷　直樹

2017 年京都大学大学院情報学研究科知能情報学専攻博士前期課程修了。自然言語処理とクラウドソーシングの研究に従事。

otani@nlp.ist.i.kyoto-u.ac.jp　　075-753-5346

京都大学大学院情報学研究科
中澤　敏明

2010 年京都大学大学院情報学研究科知能情報学専攻博士課程修了。博士（情報学）。機械翻訳の研究に従事。

京都大学大学院情報学研究科教授
黒橋　禎夫

1994 年京都大学大学院工学研究科電気工学第二専攻博士課程修了。博士（工学）。2006 年 4 月より京都大学大学院情報学研究科教授。自然言語処理、知識情報処理の研究に従事。

1　はじめに

人手評価は機械翻訳システムの性能を測る際にもっとも信頼のおける手段である。機械翻訳に限らず、文書要約や対話システムの評価のように正解と言える出力が無数に存在し、単なる表層的な判定ではその正しさが測れないような自然言語処理の幅広いタスクにおいて、人手評価は重要な手段と考えられている。

近年人手評価のコストを抑制するためにクラウドソーシングが使われるようになった。しかしクラウドソーシングの作業者は翻訳の専門家ではないため、判断はしばしば信頼性と一貫性を欠く。

作業結果の品質を上げるためには、クラウドソーシングで発注する作業は単純でなければならない。したがって、多くの先行研究は絶対評価ではなく、一対比較のような相対評価を採用している。加えて、最終成果物の誤りを減らすためには、一つの作業を複数の作業者に依頼してその判断を統合する方法が有効である。

機械翻訳の国際ワークショップが主催する機械翻訳コンペティションのために、いくつかの比較結果の統合手法が提案されている。そこでは、参加者のシステムが出力した翻訳文を人手で比較し、その結果を統合して最終的なシステムのランキングが算出されている。しかし、既存手法は次のような重要な課題を考慮できていない。

推定結果の解釈可能性：評価という目的に照らせば、評価結果にもとづいて開発者が機械翻訳システムを改善できる、あるいは評価の実施者が評価セットを点検できることが望まれる。しかし、既存手法からはシステム単位の実数スコアしか得られない。

評価者の正確さ：クラウドソーシングの作業者の中には、翻訳文の品質に関して一定の基準をもって判定を行える人もいれば、そうでない人もいる（Graham ら, 2015）。評価者個人によって異なる翻訳品質に関する正確さは重要な要素であるが、既存のモデルはそれを明示的に考慮していない。

新たに提出されたシステムの評価：これまでのアプロー

チは、提出された翻訳システム間のすべての組み合わせを考慮しており、新しいシステムが登場した場合には、そのシステムをすでに存在するシステムのすべてと比較する必要がある。

これらの課題を解決するために、我々は項目反応理論（item response theory; IRT）のモデルを利用する。この理論は学力テストやアンケートを受ける受験者の特性を評価するために用いられる。IRT モデルは高い解釈可能性を持ち、その効果は理論的・実証的研究によって示されている。例えば、テストの設問それぞれが持つ情報量を受験者の回答から推定することができる。

IRT の問題設定とのアナロジーを利用するために、我々は次のような翻訳評価の手続きを考える（図 1）。まず、あらかじめ設定したベースライン翻訳の存在を仮定する。そして、性能を求めたいシステムの翻訳を、そのベースライン翻訳と比較する。それぞれの組み合わせに対し、複数の評価者が比較結果を与える。

この一対比較は学力テストにおける設問に対応しており、評価者の正確さは設問の識別力に、一対比較でベースラインに勝利する難しさは設問の難易度に対応する。翻訳システムは学力テストの受験者と見なすことができるので、IRT モデルを使って自然にシステム（受験者）の性能（学力）を調べることができる。

さらに、あらかじめベースライン翻訳を固定することの利益として、翻訳システムが新たに提出された際に、既存手法のようにシステム間の全組み合わせを試すことなく、追加されたシステムの出力を一定数のベースライン翻訳と比較するだけで性能を推定することができる。

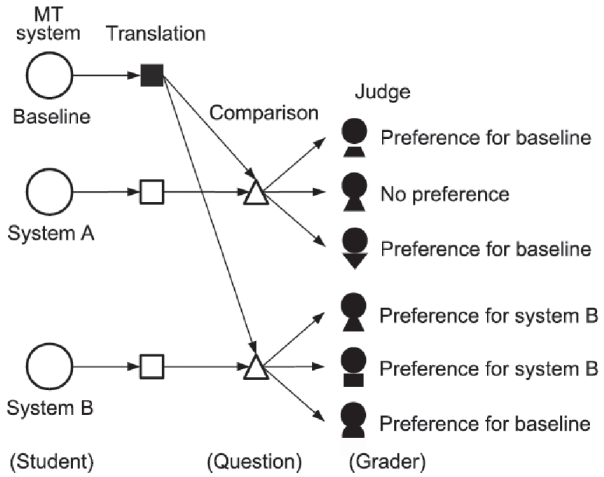


図 1　本研究の問題設定

それぞれのシステムが翻訳を生成し、評価者がベースライン翻訳と品質を比較する。

2　問題設定

翻訳システムの集合を I、入力文の集合を J、評価者の集合を K とする。人手評価を行う前に、適当なシステムを選んでベースラインとし、J に対してベースライン翻訳文を生成しておく。

翻訳システム $i \in I$ が文 $j \in J$ に対して翻訳文を生成する。そして評価者 $k \in K$ がその翻訳文をベースライン翻訳文と比較し、比較結果を報告する。

システム i が文 j に対して生成した翻訳文に関して評価者 k が与えた一対比較結果を $u_{i,j,k}$ とし、次のように値を定める。

$$u_{i,j,k}=\begin{cases}1\ (\text{ベースライン翻訳のほうが良い})\\ 2\ (\text{どちらでもない})\\ 3\ (\text{システム } i \text{ の翻訳のほうが良い})\end{cases},$$

以後、$c \in \{1,2,3\}$ で比較結果を表す。

それぞれのシステム i は潜在的な性能を表すパラメータ $\theta_i \in \mathbb{R}$ を持つとする。本研究の目標は観測された比較結果 $U=\{u_{i,j,k}\}_{i \in I, j \in J, k \in K}$ にもとづいて θ を推定することである。

3　一対比較結果の生成モデル

IRT にもとづく一対比較のモデルについて述べる。

3.1　段階反応モデルの拡張

Samejima（1968）による段階反応モデル（graded response model; GRM）をもとにして、一対比較結果の生成モデルを定義する。GRM は順序尺度を値として持つ反応を対象としている。具体的には、A+、A、B+ のようなレーティングや、学力テストにおける部分点などがその対象である。 本研究の問題設定では、一対比較結果を部分点とみなすことができる。システムの翻訳がベースライン翻訳より優れていると判断された場合、そのシステムは c=3 点を得る。もし比較結果が引き分けだった場合、システムは c=2 点を得る。そしてシステムの翻訳の負けと判定された場合、システムは c=1 点だけを得る。

$P^*_{jkc}(\theta_i)$ を文 j に対するシステム i とベースラインの翻訳に関して、評価者 k が比較結果 $\pi > c$ を報告する確率とする。その確率を次で定義する。

$$P^{*}_{jkc}(\theta_i) = \frac{1}{1+\exp\left(-a_k(\theta_i - b_{jc})\right)}$$

ただし、$P^{*}_{jk0}(\theta_i)=1$、$P^{*}_{jk3}(\theta_i)=0$ である。パラメータ a と b はそれぞれ識別力パラメータと困難度パラメータと呼ばれる。a は評価者固有の評価力を表し、b は文固有の翻訳難易度を表す。a は正の実数で、b は $b_1<b_2$ という制約を満たす。ここで b_1 はベースライン翻訳に引き分けあるいは勝つ(c>1)難しさに対応し、b_2 はベースライン翻訳に勝つ（c>2）難しさに対応する。

一対比較結果 $u_{i,j,k}$ の生成確率は上で定義した確率の差として定義される。

$$P_{jkc}(\theta_i) = P(u_{i,j,k} = c|\theta_i, b_j, a_k) = P^{*}_{jkc-1}(\theta_i) - P^{*}_{jkc}(\theta_i)$$

このモデルはもとの GRM とは異なる。GRM では a の値は項目ごとに固有な値として定義されており、それぞれの a はただひとつの項目に属す。一方で本研究のモデルは、ひとりの評価者が複数の文を評価する状況を表現するために、a を文 j とは独立な値として扱う。

3.2 パラメータの事前分布

推定値を安定して得るために、パラメータに事前分布を仮定する。θ と b には正規分布を仮定し、$\theta \sim N(0, \tau^2)$、$b_c \sim N(\mu_{bc}, \sigma^2_{bc})$ $(c=1,2)$ とする。識別力パラメータは正の値のみを取るので、対数正規分布を事前分布として仮定する。つまり、$\log(a) \sim N(\mu_a, \sigma^2_a)$ である。τ、μ、σ はハイパーパラメータである。

4 パラメータの推定

観測した一対比較結果 U に対して、以下の尤度関数を最大化するモデルのパラメータを求める。

$$L(\theta, \xi) = \log P(U;\theta, \xi).$$

本節では、$a = \{a_k\}_{k\in K}$ と $b = \{b_{j1}, b_{j2}\}_{j\in J}$ をまとめて ξ で表記する。

4.1 周辺尤度最大化

求めたいシステムの潜在パラメータ θ とその他のパラメータ ξ を同時に最適化した場合、正確な推定値が得られにくいという問題がある。そのため、まず θ に関して尤度関数を周辺化し、それを最大化する ξ を求める。周辺化した尤度関数は

$$mL(\xi) = \log P(U, \xi) = \sum_{i\in I} \log \int_{-\infty}^{\infty} P(\theta)P(U_i|\theta, \xi)d\theta + \log P(\xi),$$

である。ここで U_i はシステム i の翻訳に対する一対比較結果の集合である。積分部分はガウス・エルミート求積法（Gauss-Hermite quadrature）によって近似する。

この近似した周辺尤度関数を勾配降下法によって最適化する。このとき、バリア関数を導入することによってパラメータの不等式制約を加える。

4.2 事後分布最大化

ξ の推定値が得られたうえで、次にシステムの潜在パラメータ $\theta = \{\theta_i\}_{i\in I}$ を事後分布最大化によって求める。

最大化する目的関数は、

$$L(\theta) = \log P(U, \theta; \xi) \sum_{i\in I} (\log P(\theta_i) + \log P(U_i|\theta_i; \xi))$$

である。勾配降下法によって θ の推定値を得る。

5 実験

機械翻訳の国際ワークショップ Workshop on Statistical Machine Translation（WMT）2013 において収集された一対比較結果を用いて実験を行う。WMT2013 のデータには 10 言語対に関する結果が含まれている。詳細は Bojar ら（2013）の概要論文を参照されたい。

5.1 設定

提案手法（GRM）：a=1.7、b=（−0.5,0.5）として初期化する。θ の初期値は各システムに対して一対比較結果の値を合計したうえで、事前分布に適合するように値をスケーリングする。ハイパーパラメータは $\tau=\sqrt{2}$、$\mu_a = \log(1.7)$、$\sigma_a = 1.0$、$\mu_b = (-0.5, 0.5)$、$\sigma_b = 2.0$ とする。

比較手法：ExpectedWins（**EW**）（Bojar ら , 2013）、Hopkins と May（2013）による手法（**HM**）、そして、Baba と Kashima（2013）による二段階クラウドソーシングモデル（**TSt**）を提案手法と比較する。また、実験の正解データを生成するために使用する TrueSkill（**TS**）（Sakaguchi ら , 2014）のスコアを参考値として報告する。Sakaguchi ら（2014）の実験設定にし

たがい、HM と TS のハイパーパラメータを設定する。TSt に関しては Baba と Kashima（2013）の実験設定にしたがう。

一対比較結果：WMT2013 では 5 システムごとの部分ランキングの形で比較結果が収集されている。あらかじめこの部分ランキングを一対比較結果に変換しておく。その一対比較結果をランダムに 800、1,600、3,200、6,400 個抽出する。ただし、TS は次に入力する一対比較結果をランダムではなく能動的に選ぶ手法である。この挙動をシミュレートするため、一旦すべてのデータを受け取り、一定数に達するまで比較結果を戦略的に順次選択するという方法をとる。この点で、TS は他の手法よりも有利な環境にある。

正解データ：WMT2014 以降の公式スコア計算方法にしたがい、TS を用いて正解のシステムスコアを得る。全データから 1000 個のブートストラップサンプルを作成し、それぞれのうえで TS によって各システムの潜在スコアを計算する。この 1000 個のスコアの平均を各システムの正解スコアとする。

評価指標：正解スコアとのピアソン相関係数と nDCG（normalized discounted cumulative gain）によって推定スコアの良さを評価する。ピアソン相関係数は値の大小だけを評価する一方で、nDCG は順位を考慮する。これはシステムの順位付けを行うという今回の問題設定に適している。

5.2　結果

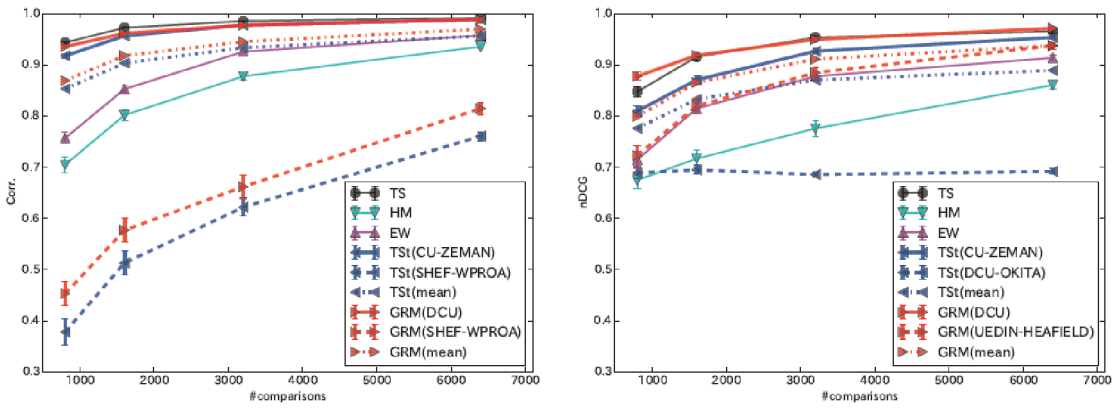
図 2 に、正解スコアと推定スコアの間の相関係数と nDCG を示す。データは WMT2013 スペイン語－英語翻訳タスクの一対比較である。ベースライン翻訳を必要とする GRM と TSt については、ベースラインとして使われたシステムをカッコ内に示している。

TS によって正解スコアが計算されていることに加え、一対比較結果を能動的に選択しているため、TS の推定スコアの相関係数と nDCG が最も高い。GRM はベースラインの設定が最も良かったケース（DCU）で TS と同等の相関係数と nDCG を達成している。一方 TSt は相関係数が最も高かったケースでも nDCG は高くなかった。つまり順位という観点では正解を再現できなかった。最も悪いケースでは GRM、TSt ともに正解と相関係数の低い値を推定したが、GRM の nDCG は高かった。このことから、GRM は正解データを十分高い精度で推定できると言える。

5.3　ベースラインの選択

ベースライン翻訳の品質が高すぎたり低すぎたりした場合、一対比較から有用な情報が得られないと予想できる。実際に図 2 の結果はベースラインシステムの選択は最終的な推定スコアに影響を与えることを示している。例えば、SHEF-WPROA の翻訳をベースラインとすると、推定されるシステムの潜在スコアは不正確だった。その理由は、SHEF-WPROA が 69.4% の一対比較で負けと判定されているからである。対象的に、DCU は 34.5% で勝ち、34.8% で負けており、その結果から他のシステムをうまく識別することができる。ゆえに、DCU をベースラインとして用いたときに最良の相関係数と nDCG を達成することができた。

では、一対比較実行前にどうやってこのような良い



(a) Correlation　　(b) nDCG

図 2　正解スコアと推定スコアの間のピアソン相関係数と nDCG

ベースラインシステムを探せば良いのだろうか。ひとつの方法は、BLEU や METEOR のような自動評価指標の結果を調べることである。自動評価の結果と一対比較統合の結果を分析したところ、自動評価の結果が全システムの平均値に近いものを選べば、比較的良い統合結果が得られることがわかった。

5.4 評価者の正確さ

評価者の正確さに対する GRM の頑健さを調べるために、不正確な評価者をシミュレートして実験を行った。実験では評価者の一部をランダムに選択し、その評価者の全判定結果を一様な確率で付け替えた。不正確な評価者の割合は 10% から 50% まで 10%ずつ増加させた。それぞれのケースで HM と EW、そして提案手法による推定スコアの精度を計算した。

表 1 に示す結果から、GRM は HM と EW よりもノイズによる悪影響が少ないことがわかる。GRM は評価者固有の正確さパラメータを持っており、不正確な評価者を検出してその影響を抑制できていると考えられる。

表 1　不正確な評価者のシミュレーション
GRM のスコアは全ベースラインの平均値である。HM と EW については GRM からの差分を示す。

ノイズ (%)	0	10	20	30	40	50
Correlation						
GRM	.929	.917	.900	.879	.849	.807
HM	+.002	−.005	−.009	−.015	−.025	−.038
EW	−.025	−.028	−.035	−.038	−.040	−.046
nDCG						
GRM	.883	.867	.847	.822	.793	.752
HM	−.024	−.130	−.137	−.144	−.152	−.168
EW	−.035	−.054	−.064	−.060	−.060	−.069

5.5 一対比較の難易度

本研究の提案手法は IRT の GRM の自然な拡張であり、IRT モデルの種々の分析方法が適用できる。標準的な分析指標である、項目の情報関数（information function）は項目（本研究では翻訳文に対応する）ごとの潜在パラメータ測定の信頼度を表す。情報関数は推定された ξ を用いて次のように定義される。

$$I_j(\theta) = -E\left[\frac{\partial^2 L(\theta;\xi)}{\partial\theta^2}\right] = \sum_{c=1}^{3}\left[\frac{\partial^2 \log P_{jkc}(\theta)}{\partial\theta^2}\right]P_{jkc}$$

$$= \sum_{c=1}^{3}\frac{[P^{*\prime}_{jkc-1}(\theta) - P^{*\prime}_{jkc}(\theta)]^2}{P^{*\prime}_{jkc-1}(\theta) - P^{*\prime}_{jkc}(\theta)}$$

ここで $P' = \partial P/\partial\theta$ である。情報関数は翻訳文固有であり、評価者からは独立であるため、すべての評価者上で一様に $a_k = 1\ (k \in K)$ とする。

表 2 にふたつの文に関する参照文と各システムの翻訳文を示す。文 1858 の情報関数は高い θ に対して高い値を持つが、文 1818 の情報関数は低い θ に対して高い値を持つ。

実際の翻訳を見ると、文 1818 に対するベースラインシステムの翻訳は比較的良いが、文 1858 に対する翻訳には “drink” や “galaxias” のような誤りが多数含まれていることがわかる。その結果、低い潜在パラメータ θ を持つシステムは文 1858 でベースラインに負けやすい。一方で低いパラメータを持つシステムでも文 1818 ではベースラインに勝つものがあり、システムの翻訳性能の識別に貢献している。情報関数は受験者の能力を効果的に測定できるように学力テストを設計する際に用いられる。同様に、機械翻訳の評価においても今回の推定結果をもとに次回の評価に含める入力文を選定することに使用できるだろう。

6 おわりに

本研究は機械翻訳評価のための人手の一対比較結果の統合に取り組んだ。本研究の狙いは次の三点にまとめられる。（1）解釈可能性の高い推定結果を得る。（2）評価者固有の正確さを考慮する。（3）新しく提出されるシステムの評価を容易にする。

これらの課題に対処するため、本研究は固定したベースライン翻訳との一対比較に着目した。一対比較の片方を予め決めておくことにより、学力テストとのアナロジーを利用して IRT の段階反応モデルを自然に適用することができる。また、すべての組み合わせを比較する場合に比べて、システムが追加された際に必要な比較数が少なく抑えられる。WMT2013 で収集された実データを用いた実験により、提案手法の頑健さと推定値の解釈可能性を確かめた。

表２　WMT2013 スペイン語ー英語翻訳タスクのデータに対する推定結果の例

参照訳は主催者によって与えられており、評価者にも提示されている。θ の推定値は 6,400 個のランダムサンプル 100 セットで得た値の平均値である。

文 1858: ベースライン翻訳に勝つことが困難		
原文		Hasta 2007 los dos telescopios Keck situados en el volcán hawaiano de Mauna Kea eran considerados los más grandes del mundo.
参照訳		Until 2007, the two Keck telescopes at the Hawaiian volcano, Mauna Kea, were the largest in the world.
DCU[baseline]		Until 2007, the two Keck telescopes located on the Hawaiian volcano Mauna of KEA were considered the largest in the world.
ONLINE-B	$\theta =$ 0.24	Until 2007 the two Keck telescopes located on the Hawaiian volcano Mauna Kea were considered the largest in the world.
UEDIN	0.12	Until 2007, the two Keck telescopes located on the Hawaiian volcano of Mauna Kea were considered the largest in the world.
CU-ZEMAN	-0.10	Until 2007, the two Keck telescope located in the volcano Mauna Kea hawaiano of were regarded as the world’s largest.
JHU	-0.12	Until 2007, the two Telescope Keck located in the Kea volcano hawaiano of Mauna were considered the world’s largest.
SHEF-WPROA	-0.92	Until 2007 the two telescope Keck located volcano hawaiano of Mauna KEA were regarded larger of world.
文 1818: ベースライン翻訳に勝つことが容易		
Source		Dependiendo de las tonalidades, algunas im · ágenes de galaxias espirales se convierten en verdaderas obras de arte.
Reference		Depending on the colouring, photographs of spiral galaxies can become genuine work of art.
DCU[baseline]		Depending on the drink, some images of galaxias galaxies become true works of art.
ONLINE-B	0.24	Depending on the shades, some images of spiral galaxies become true works of art.
UEDIN	0.12	(Same as ONLINE-B)
CU-ZEMAN	-0.10	Depending on the tonalidades, some images of spirals galaxies become true works of art.
JHU	-0.12	Depending on the tonalidades, some images of galaxies spirals become true works of art.
SHEF-WPROA	-0.92	Depending on the tonalidades, some images of galaxies spirals become real artwork.

参考文献

- Yukino Baba and Hisashi Kashima. 2013. Statistical quality estimation for general crowdsourcing tasks. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 554–562, New York, USA, August. ACM Press.
- Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT)*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

  Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2015. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28, September. (1).

  Mark Hopkins and Jonathan May. 2013. Models of Translation Competitions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1416–1424, Sofia, Bulgaria. Association for Computational Linguistics.

  Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient Elicitation of Annotations for Human Evaluation of Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT)*, pages 1–11, Baltimore, Maryland, USA.

  Fumiko Samejima. 1968. Estimation of latent ability using a response pattern of graded scores. *ETS Research Bulletin Series*, 1968(1):i–169, June.