参照文を用いない暫定的な翻訳評価と翻訳辞書作成ツールの開発

村上 浩司　　須藤 清　　新里 圭司
楽天株式会社 楽天技術研究所
{koji.murakami,kiyoshi.sudo,keiji.shinzato}@rakuten.com

1 はじめに

機械翻訳および人手翻訳のどちらにおいても翻訳結果の評価は，翻訳の品質を把握するだけでなく更に品質を向上させるために重要な手続きであり，多くの手法が提案されている [3]．翻訳の評価のためには，本来こう翻訳されるべきという参照訳 (翻訳正解文) が必要で，これに対して翻訳結果の文がどれくらい意味的もしくは表層上近いかを計測することでその品質を知ることができる．

これに対して，現実的には評価したい翻訳文に対する正解が存在しないことが多い．例えば E-commerce である楽天市場[1]は 2 億を越える商品を取り扱うマーケットプレイスであるが，出店している店舗のうち一部は，Rakuten Global Market(以下，RGM)[2]という英語のサイトにも楽天市場に出品する商品の英語説明を用意している．RGM には大凡 1,700 万の商品が約 10,000 のカテゴリに登録されている．こうしたページの多くは，英語話者や一定水準以上の日英翻訳者が翻訳に携わっているわけではなく，Google Translate[3]や BING[4]などの無料翻訳サービスを用いている場合が多く，翻訳品質はページによって大きな差がある．典型的な翻訳誤りの例を次に示す．

(1) 象印 圧力ＩＨ炊飯ジャー なべ B471-6B (楽天市場)

(2) Elephant marked pressure IH cooking a pot B471-6B (RGM)

サービスを運営する側としては，こうした英語の商品説明ページのうち，翻訳品質の低いページを特定してまとめて修正したいというニーズがある．こうしたニーズに応えるためには，何らかの尺度で翻訳の評価を行い，評価の低いものを選別する必要がある．しかしながらこれらの英語の商品説明ページには正解が存在しないため，既存の翻訳尺度を利用することができないという問題がある．

我々はこうした状況において，英語に翻訳された商品説明のタイトルを対象として（１）外部の英語情報源を利用して暫定的な翻訳の妥当性評価を行い翻訳精度の低いページを選別,（２）翻訳誤りの表現を特定し，もっとも妥当と思われる翻訳ペアを半自動的に発見,（３）簡単な作業により正しい翻訳ペアの辞書を作成，を行うためのツールを開発している．本論文ではこのツールについての詳細について述べる．

2 翻訳の評価

2.1 評価とその尺度

評価したい翻訳結果が大量にある場合，計算機によって翻訳評価が可能であれば，評価コストは低く抑えることができる．計算機による翻訳の自動評価を行うための尺度には，例えば翻訳結果と参照訳を比較して n-gram 一致率に基づいて精度の評価を行う BREU[8]，単語誤り率である WER[5]，それを元にした TER[11]，翻訳結果内の単語の並べ替えの情報を重視した RIBES[4]，表現の微妙な違いを考慮する METEOR[1] など数多く提案，利用されている．しかしながらいずれにせよ，これらの尺度を用いた翻訳評価には参照訳が必要なことには変わりない．

一方，評価を人手で行うことも考えられる．この場合，必ずしも翻訳結果に対して参照訳を用意する必要はないが，評価の基準が曖昧になる可能性があること，大量に評価する場合にはコストが高くなるなどの問題がある．

2.2 参照訳を利用しない翻訳評価

一方で，参照訳を用意せずに翻訳評価を行う試みも行われている．同じドメインの人手で翻訳された文を正例，翻訳結果文を負例として分類器を学習することで人手の翻訳に近い文に高い評価を与える手法 [2] や，対訳コーパスから IBM1 スコアを計算して利用する手法 [9]，言語横断文間含意関係の応用として翻訳評価を行う手法 [6]，

[1] www.rakuten.co.jp
[2] global.rakuten.com/en/
[3] https://translate.google.com
[4] https://bing.com/translator

[**rakuten.com:**] Home & Outdoor > Home Appliance > Small Kitchen Appliance > Rice Cookers
[**global.rakuten.com:**] Home Appliances & Small Electronics > Kitchen Appliance > Rice Cookers
[**rakuten.co.jp:**] 家電 > キッチン家電 > 炊飯器

図 1: RDC/RGM/楽天市場の間で対応するカテゴリの例

元言語に再翻訳した結果と翻訳前の元言語文から BLEU を計算する手法 [10] などがある．

我々のアプローチは [2] と同様に同じドメインの目的言語側のコーパスを用いるが，分類器による判定スコアを利用せず，より単純に翻訳結果と目的言語側の言語情報との n-gram の一致率から翻訳の妥当性評価を行い，統計情報から正しい表現の対訳ペアを求めることで対訳辞書を作成する．

3 取り組むタスク

3.1 利用するデータと外部言語資源

扱うデータは楽天市場に出品されている商品の日本語説明と，翻訳された英語説明である．英語説明をもつ商品は楽天市場全体の約 9%，1,700 万商品であり，大凡 10,000 の商品カテゴリ中に存在する．英語，日本語の両言語の説明をもつ商品は HTML 上で相互にリンクしており，翻訳された英語の品質を考慮しなければ，形式上パラレルコーパスとして利用可能である．

目的言語側の外部言語資源として，米国の E-commerce 会社である Rakuten.com Shopping(以下，RDC)[5]に登録されている英語の商品情報を利用する．このサイトには，約 6,000 万製品が 11,000 ほどのカテゴリに登録されている．RDC と RGM のカテゴリツリーの間でのカテゴリのアライメントを予め行い，対応するカテゴリ以下のサブツリー中のカテゴリに登録される製品情報を用いる．マーケットプレイスで対応するカテゴリの例を図 1 に示す．本論文では，この “Rice Cookers カテゴリ” 中の商品情報，特にタイトルの翻訳に焦点を当て説明する．以下，商品説明群のことをデータセットと呼ぶ．

RDC の “Rice Cookers” カテゴリには 2,037 商品が登録されており，それらのタイトルをすべて小文字に変換した後単語 n-gram($n = 1, 2, 3$) を予め獲得する．この結果，1,116 の unigram，2,252 の bigram，2,672 の trigram が得られた．一方，翻訳評価を行う RGM の “Rice Cookers” カテゴリからは商品のレビューの数でソートした後，上位 1,800 の商品情報を取得して利用する．同じ商品の日本語タイトル中の n-gram セット ($\boldsymbol{t}_k$) と英語タイトル中の n-gram セット ($\hat{\boldsymbol{t}}_k$) とのペアのデータセットを $D = \{\langle \boldsymbol{t}_k, \hat{\boldsymbol{t}}_k \rangle\}_{k=0}^{K}$ と定義する．英語は空白区切り，日本語は IPA 辞書により単語分割を行いそれぞれ n-gram を作成する．

[5]www.rakuten.com

3.2 翻訳の妥当性評価

人手もしくは無料翻訳サービスにより翻訳された商品情報の翻訳の妥当性評価を行う．ここでいう妥当性評価とは，我々は翻訳品質の低い商品タイトルを特定し修正するという目的から，翻訳結果全体の流暢性や文法的な正しさよりも商品が属するカテゴリや商品そのものを説明できる単語や表現で翻訳されているかを評価することである．

我々は RGM と RDC の商品カテゴリが意味的にマップされている場合，つまりマップされているカテゴリに属する商品が同じ種類の商品であると仮定できる場合に，RDC 側に登録されている商品タイトルに含まれる単語や表現を暫定的な正解表現セットとして，RGM に登録される各商品説明との間の一致率からスコアを求めることで暫定的な評価とし，類似度の低い商品群を特定する．

具体的には以下の手続きで行う．基本的に [7] で述べられている BLEU の導入と類似しており，単語 n-gram の n の変更 ($n = 1, 2, 3$) と，計算をコーパス毎ではなく各商品タイトルで行うところが異なる．評価したい商品タイトル $\{\hat{\boldsymbol{t}}_1, \ldots, \hat{\boldsymbol{t}}_K\}$ の各文 $\hat{\boldsymbol{t}}_k$ から n-gram 数を $c_n(\hat{\boldsymbol{t}}_k) = ||\hat{\boldsymbol{t}}_k|| - n + 1$ として求める．次に単語 n-gram ($n = 1, 2, 3$) 毎に正解単語 n-gram セット $\boldsymbol{U} = \{u_1, \ldots, u_M\}$ との間の n-gram 一致数 $h_n(\hat{\boldsymbol{t}}_k, \boldsymbol{U})$ を求める．ある n-gram x が $\hat{\boldsymbol{t}}_k$ に出現した頻度を表す関数 $c(\hat{\boldsymbol{t}}_k, x)$ を定義すると n-gram 一致数は次の式で計算できる．

$$h_n(\hat{\boldsymbol{t}}_k, \boldsymbol{U}) = \sum_x min(c(\hat{\boldsymbol{t}}_k, x), c(\boldsymbol{U}, x)) \quad (1)$$

これらの関数を用いて $\hat{\boldsymbol{t}}_k$ の n-gram 一致率を計算する．

$$a_n(\hat{\boldsymbol{t}}_k, \boldsymbol{U}) = \frac{h_n(\hat{\boldsymbol{t}}_k, \boldsymbol{U})}{c_n(\hat{\boldsymbol{t}}_k)} \quad (2)$$

最終的なスコアは n 毎の一致率に重みを与えた合計とする．重みは経験的に $\alpha = 7, \beta = 5, \gamma = 2$ とした．

$$Score = \alpha a_1(\hat{\boldsymbol{t}}_k, \boldsymbol{U}) + \beta a_2(\hat{\boldsymbol{t}}_k, \boldsymbol{U}) + \gamma a_3(\hat{\boldsymbol{t}}_k, \boldsymbol{U}) \quad (3)$$

このスコアを用いて商品タイトルの翻訳妥当性を評価し，作業者に提示する．

3.3 対訳辞書作成

評価スコアが低くなる大きな要因として，対象データセット内では比較的文書頻度が高いが，ドメインに

依存する固有表現や単語が誤訳になっていることが挙げられる．例 (1)，(2) 示すように “象印” というメーカ名が “elephant marked” となるのが典型的な誤訳である．こうした表現は無料の翻訳サービスではカバーしきれない語彙であり，作業者側で正しい翻訳ペア “象印:ZOJIRUSHI” を作成して対訳辞書に登録する必要がある．ツール内でこの辞書を対象データセット全体に反映させ，翻訳誤り訂正を行うことで翻訳精度の底上げが可能となる．

3.3.1 誤訳表現の特定

データセット全体の翻訳の品質を向上させるには，翻訳評価で一致しなかった n-gram のうち，頻度の高いものに注目する必要がある．“Rice Cookers” カテゴリ内で，正解単語 n-gram セットと一致しなかった単語 (1-gram) のうち，頻度の高いものは “#”，“elephant”，“messenger”，“kama” などであった．このうち，明らかに翻訳と関係ない特殊文字 (例えば装飾用の “#” など) は作業者によりストップワード辞書に登録する．しかしながら例えば “elephant” は本来，例 2 のように 2 語で出現していることから，この段階では正しい長さの誤訳とはいえない，そこでこうした単語を含む 2-gram/3-gram を対象に n-gram 内の単語間の自己相互情報量 (PMI) を計算し，値の高い n-gram も同時に作業者に提示する．2 単語間の PMI，3 単語間の PMI[12] は以下のように計算される．

$$PMI(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)} = log_2 \frac{C(x,y)\cdot N}{C(x)C(y)} \quad (4)$$

$$PMI(x,y,z) = log_2 \frac{P(x,y,z)}{P(x)P(y)P(z)+P(x)P(y,z)+P(x,y)P(z)} \quad (5)$$

これにより “elephant(頻度:315)” は，“elephant seal(頻度:221，PMI:2.508)”，“elephant marked(頻度:76，PMI:2.515)” として出現しやすいことが分かる．

3.3.2 誤訳ペアの同定と対訳ペアの生成

翻訳結果側の誤訳表現は，日本語商品説明の何からの表現を翻訳したものであるが，この時点ではその表現は不明である．日本語側の表現を特定し，更に正しい翻訳との対を作成する．データセット中の翻訳は先に述べたように無料翻訳ツール等により得られたものが多いが，人により適切に翻訳された商品もある．そのため，日本語説明中の何らかの表現 u が例えば “elephant marked” のように $\hat{u}'$ へ誤訳される場合と正しく $\hat{u}$ に翻訳される場合があり，これらの英語表現は同一商品説明には出現しないと考えられる．

そこで商品説明を，特定された誤訳表現を含むグループと含まないグループの 2 つに分け，それらの日本語商品説明中の各 n-gram について χ^2 値を計算し，誤訳表現を含むブループに特徴的に表れる n-gram と，誤訳表現とのペアを誤訳ペア候補とする．

表 1: 計算に必要なパラメータ

	$\hat{u}_i$ を含む	$\hat{u}_i$ を含むまない	
u_j を含む	a	b	
u_j を含まない	c	d	
			n

先に特定された誤訳表現の n-gram を $\hat{u}_i$ とすると，データセットはこれを含む商品説明セット $D'_{ej} = (\{\langle \boldsymbol{t}_k, \hat{\boldsymbol{t}}_k\rangle\}_{k=0}^{K} | \hat{u}_i \in \hat{\boldsymbol{t}}_k)$ と含まないセット $D''_{ej} = (\{\langle \boldsymbol{t}_k, \hat{\boldsymbol{t}}_k\rangle\}_{k=0}^{K} | \hat{u}_i \notin \hat{\boldsymbol{t}}_k)$ に分割できる．次に D'_{ej} 内の商品説明ペア中，$\boldsymbol{t}_k$ に含まれるすべての n-gram，$\boldsymbol{t}_k = (u_1, \ldots, u_J)$ について χ^2 値を計算する．計算するために必要なパラメータは表 1 のように表現でき，次の式で計算される．

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \quad (6)$$

ここで $\hat{u}_i$ を “elephant marked”，u_j を “象/印” とした場合，$a = 220, b = 1, c = 310, d = 1269$ となり，χ^2 値は第 1 位になった．作業者は提示される対訳候補の中から，最適なものを選択する．これにより，誤訳ペア “象/印:elephant marked” を得ることができた．

次に，本来 “象/印” が翻訳されるべき表現を特定し，対訳ペアを生成する．手続きそのものは誤訳 n-gram である $\hat{u}_i$ から日本語商品説明中の u_j の χ^2 値を求めたときと同様に，“象/印” を u_j として，これを含むセット D'_{je} と含まないセット D''_{je} に分割し，今度は D'_{je} 中の $\hat{u}_i$ について χ^2 値を求めれば良い．計算結果から，“ZOJIRUSHI” が上位第 5 位となった．これにより，正しい対訳ペア “象/印:ZOJIRUSHI” を生成することができる．作業者はこうした頻度の高い誤訳表現について正しい対訳ペアを生成することで，着目しているカテゴリやドメインについての無料翻訳ツールでカバーできない表現について翻訳辞書を作成することができる．

4 システム

現在開発中の翻訳辞書作成ツールのスナップショットを図 2 に示す．作業の容易さから，サーバとの通信をできるだけ行わず作業者のブラウザで動作する．システムは Bootstrap[6] フレームワーク上に構築し，ダッシュボード上で作業の切り替えを行い，データセットの読み込み，翻訳妥当性評価，対訳辞書作成を行う．追加した辞書項目は随時データセットに適用し，ストップワードや未修正の n-gram の割合を色別に表示して作業者が作業状況を把握しやすいようにしている．

[6] http://getbootstrap.com/

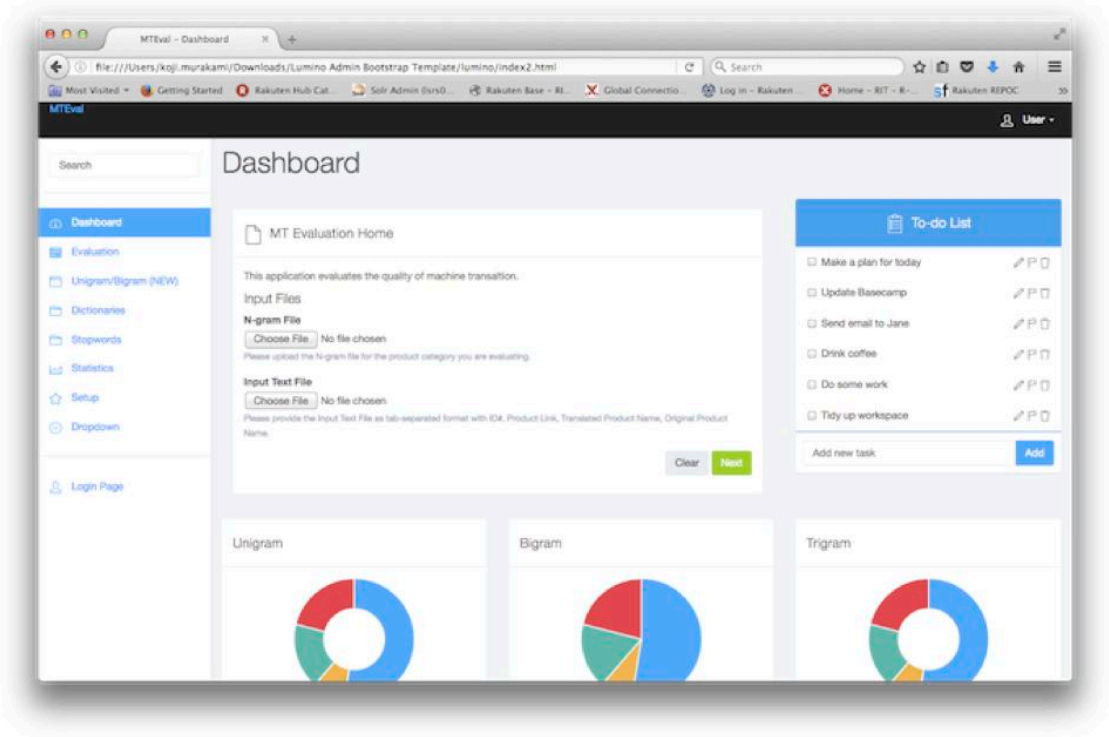

図 2: システムのスナップショット

5 おわりに

本論文では商品タイトルを対象とした翻訳妥当性評価，対訳辞書作成を行うための現在開発中のツールの詳細について述べた．正解翻訳文が存在しない状況の中で，翻訳の品質評価をするために，外部の EC サイトのカテゴリを対応付け，そのカテゴリの商品情報を利用することで擬似的な正解表現セットとすることで暫定的な翻訳妥当性評価を行えるようにした．

今後はツールの更なる開発，作業者によるシステムの主観評価や，構築した辞書を適用した商品タイトルの修正などに取り組む予定である．

参考文献

[1] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proc. of the Second Workshop on Statistical Machine Translation*, pp. 228–231, 2005.

[2] Michael Gamon, Anthony Aue, and Martine Smets. Sentence-level mt evaluation without reference translations: Beyond language modeling. In *In Proc. of EAMT 2005 Conference*, pp. 103–111, 2005.

[3] Aaron Li-Feng Han and Derek Fai Wong. Machine translation evaluation: A survey. *arXiv preprint arXiv:1605.04515*, 2016.

[4] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In *Proc. of EMNLP*, pp. 944–952, 2010.

[5] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, Vol. 10, No. 8, pp. 707–710, 1966.

[6] Yashar Mehdad, Matteo Negri, and Marcello Federico. Match without a referee:evaluating mt adequacy without reference translations. In *Proc. of 7th Workshop on Statistical Machine Translation*, pp. 171–180, 2012.

[7] Graham Neubig. 文レベルの機械翻訳評価尺度に関する調査. 情報処理学会第 212 回自然言語処理研究会 (NL-212), pp. 1–8, 2013.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *In Proc. of ACL*, pp. 311–318, 2002.

[9] Maja Popovic, David Vilar, Eleftherios Avramidis, and Aljoscha Burchardt. Evaluation without references:ibm1 scores as evaluation metrics. In *Proc. of 6th Workshop on Statistical Machine Translation*, pp. 99–103, 2011.

[10] Reinhard Rapp. The back-translation score: Automatic mt evaluation at the sentence level without reference translations. In *Proc. of the ACL-IJCNLP, Short papers*, pp. 133–136, 2009.

[11] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhou. A study of translation edit rate with targeted human annotation. In *In Proc.of Association for Machine Translation in the Americas*, pp. 223–231, 2006.

[12] Ming-Wen Wu and Keh-Yih Su. Corpus-based automatic compound extraction with mutual information and relative frequency count. In *Proc. of the ROCLING VI*, pp. 207–216, 1993.