言語横断ECサイトにおける翻訳精度改善に向けた取り組み

中澤 敏明　　塚本 浩司　　増山 毅司　　颯々野 学　　黒橋 禎夫
京都大学　　ヤフー株式会社　　京都大学
{t_nakazawa, kuro}@i.kyoto-u.ac.jp
{kotsukam, tamasuya, msassano}@yahoo-corp.jp

概 要

ヤフー株式会社と京都大学は平成 23 年度より EC サイトの中日翻訳に関する共同研究を開始した。本論文では共同研究の概要と現状を述べると共に、EC サイトドメイン対訳コーパス構築に関する考察を行う。

1　はじめに

E コマースサイトにおける総売上は年々成長しており、発展を続けている[1]。最近では国内商品のみならず、国外の商品を自分の国の言語で購入できるサービスも存在する。ヤフー株式会社が運営する「Yahoo! チャイナモール[2]」もその一つであり、中国最大級の人気ショッピングサイト「淘宝 (タオバオ)」で扱われている商品を日本人が日本語で購入できるサービスである。このようなサービスを提供するためには、原言語で書かれた各商品の商品名や説明文などを他の言語に翻訳する必要がある。しかし商品は常に入れ替わり、新しい商品が出続けるので、人手で翻訳し続けることは極めて高コストであり、機械翻訳の助けを借りる必要がある。Yahoo!チャイナモールでも中国語の説明文などはルールベース機械翻訳により日本語に翻訳されているが、商品の概要を容易に理解できるレベルには達していない。

そこでヤフー株式会社と京都大学は平成 23 年度より共同研究を開始し、EC サイトドメインの対訳コーパスを構築し、これを利用したコーパスベース機械翻訳、特に用例ベース機械翻訳 [1] による中日機械翻訳精度の向上可能性を検討している。なお本共同研究では特にファッション関連の商品を対象としている。用例ベース機械翻訳は限られたドメイン・似たドメインの翻訳において非常に有効であり、すでに自動車マニュアルの英日翻訳などにおいて既存の統計翻訳よりも高精度な翻訳を生成できることが示されている。

EC サイトドメインの対訳コーパス構築は、小説 [4]、学術論文 [3]、新聞記事 [2] などの対訳コーパス構築とは異なる問題が多く存在する。これについては 2 章で詳しく述べる。また構築コーパスを用いた翻訳実験について 3 章で述べ、最後にまとめを行う。

2　対訳コーパス作成

コーパスベースの機械翻訳手法 (統計翻訳や用例ベース翻訳) には、翻訳知識を獲得するための対訳コーパスが必須である。対訳コーパスはどのようなものでもよいわけではなく、限られたドメインにおける翻訳精度向上のためにはそのドメインの対訳コーパスを用意することが重要であり、ファッション関連の EC サイト対訳コーパスを構築することも共同研究の目標の一つである。

本共同研究では中国語の商品の Web ページから中国語の文を抽出し、これを翻訳会社に人手で翻訳してもらうことにより対訳コーパスの構築を行っている。本説では中国語を日本語に訳す際に問題となる点と注意すべき点、さらに良質な対訳コーパス構築に向けた翻訳会社とのコミュニケーションの取り方についての知見を述べる。

2.1　中 → 日翻訳におけるポイント

良質な中日ファッション関連 EC サイト対訳コーパスを構築するためには、専門用語や語彙の曖昧性、さらには各国の文化の違いにも注意を払う必要がある。以下ではこれまでの対訳コーパス構築において問題となった例を示す。

ファッション特有の表現

普段から使われる語彙であっても、ファッションドメインにおいて特別な意味を持つものが存在する。例

[1] http://www.census.gov/retail/mrts/www/data/pdf/ec_current.pdf
[2] http://chinamall.yahoo.co.jp/

えば “不规则的下摆” における “不规则” はそのまま “不規則” という意味も持つが、ここではそう訳しても意味をなさず、“波打った” もしくは “フレアの” と訳すべきである (“下摆” は “裾” という意味)。

同様に “木耳” はキクラゲという意味もあるが、ファッションドメインでは別の意味である “フリル” と訳すべきである。また “森女” という語は普段は使われないが、ここではそのまま “森女” と訳すよりも “森ガール” と訳す方がユーザーの理解の助けになる。

EC サイト特有の表現

ファッションだけではなく、EC サイト全体に特異的に出現する表現もある。例えば “秒殺 (すぐに売り切れる)” や “淘金币 (大安売り)” などである。また多くの EC サイトには出品者を評価するシステムがあり、淘宝においてもダイヤモンド (钻)、シルバー王冠 (皇冠)、ゴールド王冠 (金冠、金皇冠) のマークとその個数によって出品者の評価をしており、正しい翻訳を行うためにはこのような背景を知っておく必要がある。さらに EC サイトでは商品のレビューを記載している場合があり、多くの場合レビューを投稿したユーザーの ID が表示されている。このユーザー ID を翻訳することは無意味であるばかりか、翻訳してしまうと逆に誤りとなってしまうこともあり、翻訳するべきでない部分の同定も重要である。

文化の違いによる表現の差異

中国語では直接的な表現が好まれるのに対し、日本語では婉曲表現が多用される。これが EC サイトにも当てはまり、例えば日本語で言う “大きめサイズのレディース” は中国語では “胖女人 (太った女性)” と表現される。しかしこの中国語を直訳してしまうと、日本人ユーザーに不快感を与えかねない。

これと関係した事例として、中国語の (英語でも同様であるが) 商品説明文には “我们 (我々)“ や “您 (あなた様)“ といった表現が頻繁に用いられる。これらに該当する日本語の表現は “当店” や “お客様” とするのが妥当であり、注意を要する。

また “ようつべ” などのように特定のドメインでしか使われない表現は中国語にも存在する。ファッション EC サイトにおいては MM や GG といった表現がよく見られるが、実はこれらはそれぞれ女性と男性を意味する。これらは一種の acronym であり、女性を表す妹妹の発音が MeiMei であり、男性を表す哥哥の発音が GeGe であることから来ている。これらを日本語文において MM や GG としても日本人ユーザーには意図は伝わらない。

日本語では不自然、不適切な複合名詞

漢字という同じ文字種を共有している中国語と日本語においては、中国語の複合名詞そのままでも日本語として通じるものも少なくない。このため中日翻訳の際に複合名詞を安易に残してしまいがちであるが、日本語では不自然であったり原文と意味が異なる場合もある。例えば “特别強調” はそのままでも意味が分からないこともないが、自然な日本語に訳すならば “注意事項” とするべきである。また “着用効果図” は “着用することによる効果を示した図” ではなく、正しくは “着用時の写真” という意味であり、複合名詞をそのまま残してしまうと誤訳となる。

専門用語・固有名詞

専門用語や固有名詞は翻訳を考える上で常に問題となるが、EC サイトの翻訳でも例に漏れず大きな問題である。商品名については、“磨毛 (フリース)” や “风衣 (ウインドブレーカー)” など日本にも該当する訳語あるような一般的なものなら問題ないが、“ヒートテック” などのようなものは他の言語には対応する表現がない。会社名なども、その国の言語での名前しか持たない場合がある、これらの扱いをどうするかはあらかじめ決めておく必要がある。本共同研究では日本語に該当する表現がある固有名詞は日本語に訳し、そうでないものは中国語のまま残すことにしている。

カンマでつながる中国語

中国語はカンマを用いて文を長く続けるという特徴がある。日本語にする場合は、カンマの前後のブロックに強い依存関係が存在しないならば、複数文に分割した方が理解しやすい場合がある。例えば図 1 の例では || の位置で 3 つの文に分けると理解しやすい。

2.2 翻訳会社との意思疎通

前章で述べたような中日翻訳における注意点を翻訳会社に的確に伝え、構築する対訳コーパスに反映するためには、翻訳会社と綿密にコミュニケーションを取ることはもちろんのこと、翻訳のガイドラインを定めて文書化することも重要である。対訳コーパス構築のための翻訳は一般的な文書の翻訳とは異なり、独特の要求がある。主なものとしては**意訳をしない**ことや**訳抜け、過剰訳の禁止**、**文の 1 対 1 翻訳**がある。過剰訳は括弧くくりで専門用語などの説明を行ったりすることを指す。また 1 対 1 翻訳は、現状の多くの機械翻訳システムが仮定していることによる制約である。本共

中: 上面的刺绣和亮片均为原厂工人原厂设备精心缝制，|| 挑剔的姐妹们在看到货品之后会发现绝对可以和专柜货品比肩，而且绣工精细清晰，|| 精棉质地，密度高，手感好，穿着舒适，质量超好。

日: 上の刺繍とスパンコールは、全てオリジナル工場の作業員とオリジナル工場の日設備で心を込めて作成したものです、|| あら捜しをするお客様も、この商品を見れば、専門店の商品と匹敵し、作りが精細で、はっきりしたものだと思われるはずです、|| 精綿生地であり、密度が高く、手触りも良く、着用したえ快適で、品質もとても良いです。

図 1: カンマでつなげられた長い中国語文と翻訳会社による翻訳 (|| の位置で分割するのが妥当)

同研究ではさらに**最適な文単位への分割**も要求している。これは以下のような背景から来るものである。

- 原文である中国語文は Web ページから抽出したものであるため、文区切りに誤りがある
- 1 対 1 翻訳の実現のため (前述のように中国語にはカンマ区切りで長い文が存在する)
- 長い文は文の解析や翻訳などの処理の障害となる

この他**セクションの文体の違いを意識した翻訳**も要求している。EC サイトの各商品ページには大きく分けて商品のタイトル、サイズや色などの商品の属性、商品の説明文の 3 つのセクションがある。日本語にした場合にタイトルは文末が体言、属性は名詞や数値の羅列など、各セクションごとに文体に特徴があると考えられるため、これを考慮することを要求している。例えば “到货！ ” という表現がタイトルにある場合には “入荷！ ” と訳し、商品説明文にある場合には “入荷しました！ ” と訳すという具合である。

しかしガイドラインを定めて翻訳のルールを共有しても、誤った翻訳結果が散見されることがある。これまで再チェック時に発見された翻訳の誤り例を表 1 に示す。また図 1 に示したように、読点を用いて無理矢理日本語を 1 文として翻訳している例もあった。そもそも翻訳会社では多くの翻訳スタッフを抱えており、全てのスタッフに特別なガイドラインを徹底させることは本質的に難しく、このような翻訳誤りがある程度生じてしまうことは不可避である。

このような翻訳誤りを極力減らし、対訳コーパスの質を維持するために、本共同研究では中国語のわかる第三者の日本人による、翻訳品質抜き取り調査を行っている。この調査で翻訳の質の悪い文や翻訳が誤っている文を翻訳会社にフィードバックすることにより誤りを減らすことが可能であり、また翻訳会社からの再フィードバックにより、ガイドラインについて不明な点や曖昧な点などの不備を修正することや、新たに発覚した現象についての対処の追加なども可能である。

3 構築したコーパスでの翻訳実験

本共同研究はまだ途中段階であるが、現時点での我々のシステムの位置付けの確認やコーパスの量の評価を行うため、構築した対訳コーパスを用いた翻訳実験を行った。今回の実験では構築した対訳文のうちの一部、約 60 万対訳文 (中国語単語数約 360 万語、日本語単語数約 500 万語) を用いた。なお中国語の形態素解析には京都大学で開発している解析器を用い、依存構造解析には ALAGIN Forum よりオープンソースで公開されている CNP[3] を利用した。日本語の形態素解析器、依存構造解析器にはそれぞれ JUMAN7.0[4] と KNP4.0[5] を利用した。

テストデータは、2011 年 8 月 7 日時点で Yahoo!チャイナモールで購入可能であったファッション関連商品からランダムに 10 商品を選び、ここから抽出した中国語文 702 文、5448 語を用いた。さらに中国語のわかる日本人にこれらの中国語文を人手で日本語に翻訳してもらい、正解の翻訳とした。翻訳精度は BLEU および RIBES で評価した。評価結果を表 2 と表 3 に示す。比較として 2011 年 8 月 7 日時点で Yahoo!チャイナモールに表示されていた日本語文、Yahoo 翻訳[6]による翻訳および Google 翻訳[7]による翻訳と、EBMT でトレーニングコーパスサイズを 1/2、1/4、1/8 にした場合も評価した。また Yahoo!チャイナモールの商品ページは大きく分けて商品名 (Title)、属性値 (Feature)、商品説明 (Description) の 3 つのフィールドからなるため、それぞれのフィールドごとの精度を示した。さらにトレーニング文に含まれない語 (OOV) の割合も示した。

結果を見ると、現時点での EBMT の精度は Yahoo 翻訳と Google 翻訳よりはよいが、Yahoo!チャイナモールの翻訳はこれら 3 つと比べてかなり良いということがわかる。これは Yahoo!チャイナモールの翻訳を生成しているルールベースの翻訳システムが EC サイト用にかなりチューニングされており、また専門の対訳辞書を利用していることが大きな原因であると考えられる。しかし、現在の我々の EBMT には、解決が容易な小さな問題がいくつか残っており、それらを修正することで精度は大きく改善されることを期待している。

またトレーニングコーパスサイズを増やすと次第に

[3] http://alaginrc.nict.go.jp/cnp/index.html
[4] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[5] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
[6] http://honyaku.yahoo.co.jp/
[7] http://translate.google.co.jp/

誤り種類	中国語原文	翻訳会社による訳	正しい訳
訳し忘れ	看看有没有其他合适的商品	看看有没有その他合適的商品	他に良いものがないかご覧ください
誤訳	加湿器功能:	除湿器の機能：	加湿器の機能：
誤訳	买家秀身上穿的是两件， 一口价是一件的价格!	お客様ショーの体に着ているのは 2 点、 ワンプライスは一枚の値段です！	モデルが着ているものは 2 着で、 価格は 1 着の値段です！
不要字挿入	不要随便拍下一种	随意にに 1 種類だけ注文するのではなく	随意に 1 種類だけ注文するのではなく
中国字残り	精神焕发之效果	元气あふれるという効果があります	元気があふれるという効果があります

表 1: 翻訳会社による翻訳に含まれていた誤り例

	Title	Feature	Description	OOV
EBMT	0.00	26.46	12.90	4.35%
EBMT(1/2)	0.00	26.08	11.65	5.10%
EBMT(1/4)	0.00	22.09	11.61	6.20%
EBMT(1/8)	0.00	21.78	11.35	8.52%
チャイナモール	37.92	39.68	21.47	
Yahoo 翻訳	0.00	16.97	9.41	
Google 翻訳	0.00	21.46	11.56	
文数	13	258	431	

表 2: BLEU による翻訳精度比較

	Title	Feature	Description
EBMT	60.58	54.51	62.50
EBMT(1/2)	56.71	55.30	62.46
EBMT(1/4)	58.45	54.40	60.54
EBMT(1/8)	55.43	50.50	59.87
チャイナモール	79.58	63.27	75.06
Yahoo 翻訳	49.40	44.74	63.80
Google 翻訳	44.49	41.31	52.04
文数	13	258	431

表 3: RIBES による翻訳精度比較

精度が上がっていることがわかるが、OOV の割合は全てのトレーニングコーパスを使っても 5%弱あり、OOV の割合も翻訳精度に影響していることがわかる。これはトレーニングコーパスのサイズがまだ小さいこともあるが、中国語の未知語に対する単語分割が不安定であり、実際の単語異なり数よりも多くなってしまっているという問題もある。

4 結論

本論文では共同研究の概要と現状を述べると共に、対訳コーパス構築に関する知見を述べた。高品質な中日 EC サイト対訳コーパス構築のためには、第三者による品質チェックを行うことが非常に有効であるが、もちろん全ての文をチェックすることは難しい。訳し忘れや中国語の文字が日本語文に残っている問題などは機械による自動チェックが可能であろうが、中国語にしかない固有名詞は日本語文に残っているのが正しいため、文字コードなどによる単純なチェックでは不十分である。これを解決するためには、日中の共通漢字情報を使うことや、単語アライメントにおいて対応の確率が低い部分を検出することなどが考えられる。

またその他の課題としては、翻訳において html タグや顔文字、記号などの本質的ではない部分を適切に扱うことがある。用例ベース機械翻訳では両言語の文を構文解析しており、これらの文字は構文解析誤りの要因となっている。同様に専門用語なども単語分割・構文解析誤り、翻訳誤りの大きな原因の一つである。専門用語対訳辞書があれば翻訳においても利用可能であるが、単言語の専門用語辞書だけでも各言語の解析精度を向上することが期待される。このような専門用語辞書を既存の手法などを用いて構築することも今後の課題である。

参考文献

[1] Toshiaki Nakazawa and Sadao Kurohashi. EBMT system of KYOTO team in PatentMT task at NTCIR-9. In *In Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-9)*, pages 657–660, 2011.

[2] Yujie Zhang, Kiyotaka Uchimoto, Qing Ma, and Hitoshi Isahara. Building an annotated Japanese-Chinese parallel corpus - a part of nict multilingual corpora. In *Proceedings of 2nd International Joint Conference on Natural Language Processing*, pages 85–90, 2005.

[3] 菊池 俊一, 青木 雅子, 井上 聡子, and 蒋. 科学技術分野の日中対訳コーパス・日中専門用語の集積と公開. In **第 7回情報プロフェッショナルシンポジウム** *(INFOPRO2010)*, 2010.

[4] 曹 大峰, 中野 洋, 徐 一平, and 隈井 裕之. 中日対訳コーパスの作成状況と今後の課題. **情報処理学会研究報告. 自然言語処理研究会報告**, 99(95):1–8, 1999-11-25.