対訳表現を手がかりとした用例の選択手法の提案

大山 鉄郎†
筑波大学 情報学群 知識情報・図書館学類†
s1013134@u.tsukuba.ac.jp†

関 洋平‡
筑波大学大学院 図書館情報メディア研究系‡
yohei@slis.tsukuba.ac.jp‡

1　はじめに

機械翻訳は主な翻訳方式として，用例を参考にして訳文を生成する用例翻訳 [4, 8, 7]，文法規則などを用いるルールベース翻訳，統計的な確率を利用する統計翻訳がある。本研究では，文を流暢に翻訳できるという特徴を持つ，用例翻訳を対象として，用例の選択手法の提案を行う。

用例翻訳では，対訳コーパスを利用して，入力文に対して適切な用例を選択して用いることで翻訳が行われる。このとき，翻訳に適している用例を選択するための手がかりとして，入力文と用例との文字列や構文情報の一致度，あるいは，文中に含まれる語の概念の類似度が用いられる。

例として，入力文 “出版社について教えて下さい。” の用例選択の問題点について以下に示す。対訳コーパスから用例を検索すると，“教えて下さい。” の文字列の一致により，以下の 3 つの候補が得られる。

1. ステップを**教えて下さい。**
   — Teach me the steps, please.
2. ルールを**教えて下さい。**
   — Please tell me the rules.
3. 連絡先を**教えて下さい。**
   — Give me the phone number.

このとき，“教えて下さい。” の対訳表現候補として，“Teach me 〜 please”，“Please tell me 〜”，“Give me 〜” の 3 つが得られる。

複数の候補からの絞り込みには，用言にかかる目的語の概念の類似度や，“教えて下さい。” という語に対して，どの訳語を選択しやすいかという，翻訳確率が用いられる。しかし，この例では，“出版社” に対して “ステップ”，“ルール”，“連絡先” の 3 語とも概念的に遠すぎて比較が難しい。また，翻訳確率は 3 種類の表現が 1 回ずつ出ているため，3 つの用例とも同じ値となり比較できない。

本研究では，それらの問題に対して，“教えて下さい” の対訳表現候補を手がかりとすることで，適切な用例が選択できるような翻訳手法の提案を行い，その有効性を検証する。

2　関連研究

代表的な日英用例翻訳手法を，用例選択の手がかりを中心に紹介する。隅田 [7] は，意味距離と編集距離とを組み合わせて，入力文に対して 1 つの用例を選択して翻訳を行った。この手法は，シンプルな処理で多言語適用性に優れるが，入力文に対して 1 つの用例しか翻訳に使用できない。

荒牧ら [8] や中澤ら [4] は，文節単位で用例を選択している。用例の選択は，荒牧ら [8] が翻訳確率を利用しているのに対して，中澤ら [4] は入力文と一致する文節数を重視し，その他に翻訳確率などを組み合わせて用いている。これらの手法は，入力文に対して複数の用例を用いた柔軟な翻訳ができるが，入力文と表層表現が一致した用例以外は用例の選択処理に含まれないという問題がある。

3　対訳表現を手がかりとした用例の選択手法

本研究では，入力文中の文節と一致した用例をもとに，その用例の対訳表現と一致する用例を利用して，翻訳に使用する用例を決定する。

提案手法は 3 つの処理で構成されている。以下，それぞれの処理について説明する。

3.1　候補となる用例の検索

用例の検索の前に，あらかじめ対訳コーパス内の用例を文節ごとにアライメントしておく。文節の分割には構文解析器 CaboCha [9] を，アライメントには GIZA++[5] を使用している。候補となる用例の検索は，以下の手順で行う。

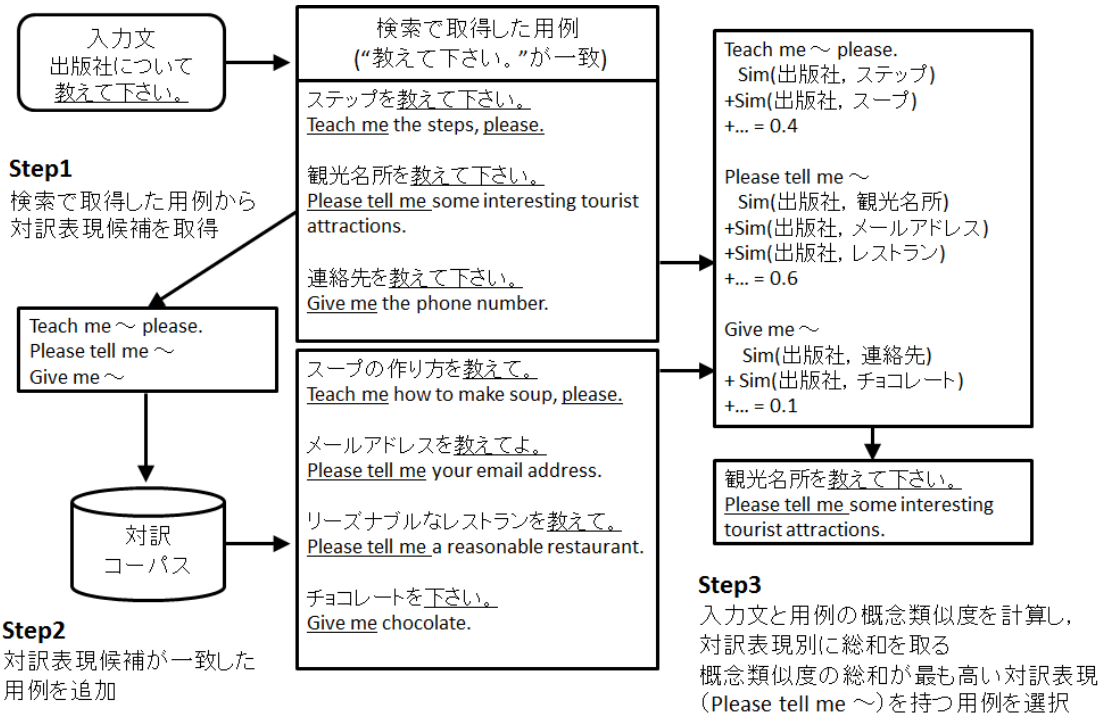

図 1: 対訳表現候補を手がかりとした用例選択手法

1. 入力文を文節に分解して，文節ごとにその文節を含む用例をすべて検索する。
2. 検索した用例を，入力文と一致する文節数でランク付けし，上位のもので用例集合を絞り込む。
3. 文節の一致数で絞り込まれた用例集合のうち，入力文との文節単位の編集距離で更に絞り込み，上位 N 件の用例を翻訳候補として取得する。

3.2 対訳表現を手がかりとした用例選択

対訳表現候補を用いて，入力文の各文節の翻訳に使用する用例を選択する。提案手法の概要を図 1 に示す。

図 1 では，入力文の一部である “教えて下さい。” について，翻訳に使用する用例を選択する処理の流れを示している。

Step1: 対訳表現候補の抽出

用例の検索で取得した用例集合から，現在，処理対象としている文節（図 1 の例では “教えてください。”）と，アライメントが取られている対訳表現を，対訳表現候補として抽出する。図 1 では，“Teach me ～ please.”，“Please tell me ～”，“Give me ～” の 3 つの表現が抽出できている。

Step2: 対訳表現候補を手がかりとした用例の追加

Step1 で得た対訳表現候補を用いて，再度，対訳コーパスからの検索を行う。図 1 では，“Teach me ～ please.” や “Please tell me ～” の検索結果として，“スープの作り方を教えて。” や “メールアドレスを教えてよ。” などの表現が追加できている。これらの表現は，Step1 で抽出した対訳表現で検索されているため，入力文の原言語側の表現（“教えてください。”）とは表層的に一致しない表現も追加できる。

Step3: 最適な対訳表現を持つ用例の選択

Step2 で追加された用例も含めた用例集合に対して，入力文の対象文節に係っている名詞と，用例の対象文節に係っている名詞との概念類似度を，日本語 WordNet [2] を用いて計算する。次に，計算した概念類似度について，Step1 で抽出した対訳表現候補別に総和を取る。図 1 では，“Teach me ～ please.”，“Please tell me ～”，“Give me ～” の表現別に概念類似度の総和を計算している。最後に，概念類似度の総和が最も高い対訳表現候補をその文節の対訳表現として，それを含む用例を選択する。図 1 では，“Please tell me ～” を，文節 “教えて下さい。” の最適な対訳表現として，それを含む用例を選択している。

3.3 翻訳処理

翻訳に使用する対訳表現が用例のどの位置にあるかという情報をもとに，選択した用例の対訳表現を組み合わせ，出力文を生成する。この際，用例から対訳表現が得られなかった文節に関しては，辞書を用いて対訳表現を獲得する。

4 実験：翻訳精度の評価

提案した用例選択手法を用いて翻訳を行い，$BLEU$[6] と WER[3] による評価値を計算する。評価値は，ベースラインと比較することにより，提案手法の有効性について検証した。

4.1 実験データ

実験データは，高度言語情報融合フォーラム ALAGIN [1]で公開されている日英翻訳エンジン学習・評価用対訳コーパス [10] から，用例として 19,972 文，テストデータとして A(506 文)，B(500 文)，C(506 文) の 3 セットを使用した。辞書には EDICT [1] を用いている。

4.2 ベースライン

ベースラインには，提案した用例選択手法のうち，3.2 節の対訳表現候補を手がかりとした用例選択での処理を，以下のように置き換えたものを用いる。

- 日本語 WordNet [2] を用いて，入力文の対象文節に係っている名詞と，用例候補それぞれの対象文節に係っている名詞との概念類似度を求め，最も概念類似度が高い用例を採用する。

ベースラインでは，同じ概念類似度の用例が複数存在することがあるが，この場合は複数の用例から 1 つをランダムで選択する。

4.3 実験結果

実験結果を表 1 に示す。提案手法はベースラインと比較すると，平均で $BLEU$ は 0.02 ポイントの精度向上，WER は 1.1%の誤り率の削減が見られた。

5 提案手法によって改善された例

表 1 により，提案手法は日本語 WordNet [2] を用いたシンプルな用例選択と比較して，若干の精度向上が見られた。これにより，提案手法である，対訳表現候補を手がかりとした用例選択手法は有効であるといえる。

提案手法によって精度が向上した理由として，以下の 2 点があげられる。

1. 入力文と表層表現が異なる用例が概念類似度の計算に含まれることで，**適切な用例が選択されやすくなった。**
2. 用例候補に含まれているが，原言語側の文節と目的言語側の**単語のアライメントが誤っている用例が選択されにくくなった。**

提案手法を用いた翻訳について，ベースラインから改善された例を，以下に示す。

入力文: メニューを下さい。
参照訳: I'd like a menu, please.
用例 A: **メニューを**見せて下さい。
Can we see a menu?
用例 B-1: 着き次第手紙を**下さい。**
Please write a letter as soon as you arrive.
用例 B-2: アイロンを**下さい。**
I'd like an iron, please.

入力文 “メニューを下さい。” は “メニューを” と “下さい。” の 2 つの文節から構成されている。“メニューを” の対訳表現はベースライン，提案手法とも用例 A を選択し，対訳表現 “a menu” が得られた。一方で，“下さい。” の対訳表現は，ベースラインが用例 B-1 を，提案手法が用例 B-2 を選択したため，最終的な出力文は

ベースライン: Please write a menu.
提案手法: I'd like a menu, please.

となり，提案手法によって出力文が改善されていることがわかる。

ベースラインでは，入力文の “メニュー” と用例 B-1 の “手紙”，用例 B-2 の “アイロン” の概念類似度を比較し，より高い値を出した用例 B-1 を選択した。

提案手法では，まず，“Please write 〜” と “I'd like 〜 please.” と一致する用例を用例集合に追加してから，概念類似度を計算した。“Please write 〜” と “I'd like 〜 please.” で追加できた用例は “I'd like 〜 please.” の方が多く，また文の概念類似度は平均して，“I'd like 〜 please.” の表現を持つ用例の方が良い値だった。結果として，概念類似度の総和は “Please write 〜” より “I'd like 〜 please.” の方が高い値となり，正しい

[1]http://alaginrc.nict.go.jp/

表 1: 翻訳精度

評価尺度	$BLEU$[6]			WER[3]		
テストデータ	A	B	C	A	B	C
ベースライン	0.2834	0.2817	0.2124	0.6214	0.6396	0.6814
提案手法	**0.3104**	**0.3057**	**0.2220**	**0.6036**	**0.6264**	**0.6788**

対訳表現である，用例 B-2 の “I’d like 〜 please.” が選択された。このように，用例候補の検索で得られた対訳表現で用例を追加することで，より多くの用例から適切な用例を判断できるようになり，意味的に近く，よく用いられる対訳表現が選択されやすくなる。

また，提案手法はアライメントが誤っている用例は選択されにくくなっている。アライメントが誤っている用例の対訳表現は，対訳コーパスに出現する数が少ない傾向がある。そのため，アライメントが誤った用例の対訳表現で用例を検索しても，追加される用例は少なく，概念類似度の総和が小さくなる傾向があるため，その対訳表現を持つ用例が選択されにくくなる。

例として，“御願いします。” という文節に対して，アライメントを誤った用例 C を検索した場合を説明する。

入力文: ボーイさん、**御願いします。**
参照訳: Waiter please.
用例 C: コーチを**御願いします。**
Coach, please.

用例 C は “御願いします。” の対訳表現として “Coach” という誤ったアライメントがとられている。次に，対訳表現である “Coach” を用いて，用例を検索し直すと，以下の用例 D のみが検索される。

用例 D: 普通車ですか、プルマン車ですか。
Coach or Pullman?

通常，対訳表現候補を用いた検索では，複数の用例を追加できるが，アライメントが誤っている対訳表現候補で検索すると，追加できる用例が少ない傾向がある。また，個々の概念類似度は小さくなる傾向がある。例では，用例の対象文節に係っている名詞がないため，概念類似度は 0 として計算される。結果として，アライメントが誤っている用例は選択されることが少なくなる。

6 おわりに

本研究では，対訳表現を手がかりとした用例選択手法を提案，評価を行った。提案手法は実験により，シンプルな概念類似度を用いた手法と比較して，$BLEU$ は 0.02 ポイントの精度向上，WER は 1.1%の誤り率の削減が見られた。提案手法は，概念類似度で適切な用例が選択できない場合に，有効に機能することを示した。

今後の課題として，入力文か用例のいずれかが多義性を持つ場合，不適切な用例を過剰に追加してしまうことがあるため，曖昧性解消処理の追加等について，検討を予定している。

参考文献

[1] EDICT. http://www.csse.monash.edu.au/ jwb/edict.html, (accessed 2012-01-30).

[2] Kow Kuroda, Francis Bond, and Kentaro Torisawa. Why Wikipedia Needs to Make Friends with WordNet. In *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010)*, pp. 9–16, Mumbai, India, 2010.

[3] Gregor Leusch, Nicola Ueffing, Hermann Ney, and Lehrstuhl Fur Informatik. A Novel String-to-string Distance Measure with Applications to Machine Translation Evaluation. In *Proceedings of Machine Translation Summit IX (MT Summit IX)*, pp. 240–247, New Orleans, USA, 2003.

[4] Toshiaki Nakazawa and Sadao Kurohashi. Fully Syntactic EBMT System of KYOTO Team in NTCIR-8. In *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-8)*, pp. 403–410, Tokyo, Japan, 2010.

[5] Franz Josef Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51, 2003.

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU:A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL2002)*, pp. 311–318, Philadelphia, USA, 2002.

[7] Eiichiro Sumita. Example-based Machine Translation Using DP-matching between Word Sequences. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL2001) Workshop on DDMT*, pp. 1–8, Toulouse, France, 2001.

[8] 荒牧英治, 黒橋禎夫, 柏岡秀紀, 加藤直人. 用例ベース翻訳の確率的モデル化. 情報処理学会論文誌, Vol. 13, No. 3, pp. 3–19, 2006.

[9] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.

[10] 日英翻訳エンジン学習・評価用対訳コーパス. http://alaginrc.nict.go.jp/images/documents/MTEVAL_ALAGIN_V1_README.pdf, (accessed 2012-01-28).