自動コーパス生成とユーザフィードバックによる機械翻訳

藤原 菜々美† 今出 昌宏† 山内 真樹†‡ 内山 将夫‡ 隅田 英一郎‡
†パナソニック株式会社 先端研究本部 インタラクティブ AI 研究部
‡情報通信研究機構 先進的翻訳技術研究室

{fujiwara.nanami, imade.masahiro, yamauchi.masaki} @ jp.panasonic.com,
{yamauchi, mutiyama, eiichiro.sumita} @ nict.go.jp

1 はじめに

訪日外国人の急増・2020年東京オリンピックに向けて，多言語翻訳機への期待が高まっている．日本国内においても，観光案内などで実証実験が始まっている[1]．現在実用化段階にある翻訳機の一つとして，統計的機械翻訳(SMT: Statistical Machine Translation)[2]が挙げられる．SMT は予め用意された対訳コーパス(言語間で同じ意味を持つ対文)から，翻訳に必要なモデルを統計的に学習し，そのモデルを通じて翻訳を行う．翻訳性能と対訳コーパスの量には相関があると言われており，質の良い大量の対訳コーパスを学習させるほど，翻訳性能が向上する．このように，質の良い対訳コーパスを数多く集めることが翻訳性能向上の鍵であるが，対訳コーパスの収集は高コストである．

翻訳可能なドメイン(領域)を増やすには，ドメインごとの翻訳機の構築が望ましい．しかし，新規ドメインの対訳コーパス収集は一般に困難であり，ドメインによらず1,000文オーダーと言われている．一方で，翻訳機の構築に最低限必要な対訳コーパスは約10万文であり，少量の対訳コーパスを用いた翻訳機の構築には，このギャップを埋めることが必要不可欠である．

これに対し我々は，少量の対訳コーパスからの翻訳機の構築を目的とし，十分量の対訳コーパスを自動的に獲得すべく，自動対訳コーパス生成手法(ACG: Automatic Corpora Generation)を開発している[3] [4] [5]．これは，種となる対訳コーパスを元に，類似候補文を自動生成し，その中から良質なコーパスを識別する枠組みである．

本稿では，類似候補文生成，候補識別を適用した統計的機械翻訳に対し，特に利用者からのフィードバックを活用し，識別器を順次更新することでの，翻訳機の性能向上手法の検討とその評価について報告する．

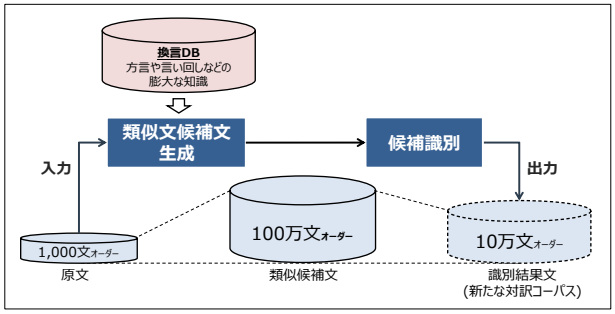

Fig. 1 Automatic corpora generation

2 自動コーパス生成技術：ACG

我々が開発中の ACG の構成概要図を Fig. 1 に示す．

ACG は，入力された対訳コーパスを元に，「類似候補文生成」器と「候補識別」器により，多量の対訳コーパス(識別結果文)を生成する．

類する先行研究としては，WordNet から言い換えに適した候補を選択し，対訳コーパスの拡張を行う手法[6]や，置換えルールでのコーパス拡張手法[7]などが挙げられる．翻訳性能を対コーパス数比で換算した場合において，先行研究では1.16～1.45倍程度であるのに対し，ACG は約3.20倍以上の効果を得ている[3]．

2.1 類似候補文生成

我々の提案手法における「類似候補文生成」器では，言い換え表現のデータベースを言語資源(WordNet [8], PPDB[9] ，内容語換言辞書[10]等)，及び手作業から構築し(換言データベース)，入力文に対して適用することで類似候補文を得る．

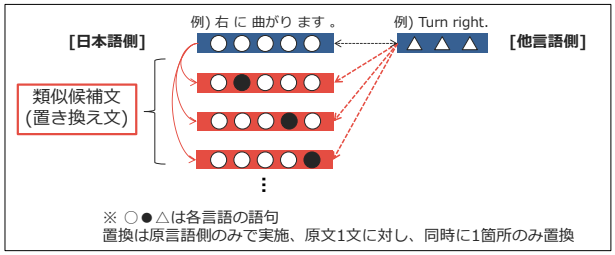

Fig. 2 Original and candidate corpora

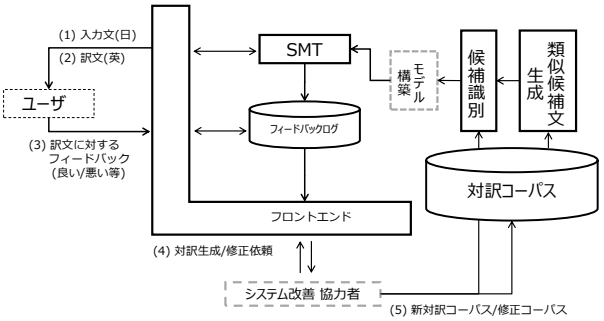

Fig. 3: Feedback system

類似候補文の生成模式図を Fig. 2に示す．原文(ここでは日本語文)1文に含まれる語句・文節に対して同時に1箇所の置換えを行う．生成された類似候補文の中には，文としての品質が必ずしも高く無く，意味的・文法的に破綻した文も生成される可能性がある．これは，対訳コーパスの想定ドメインが換言データベースのエントリと必ずしも合致しないことや，エントリ自身のノイズ等に起因する．

次段の「候補識別」器では，このような破綻文を除外し，対訳コーパスに適切な文を識別する．

2.2 候補識別

提案手法における「候補識別」器では，類似候補文に対して識別器を適用し，“良い文”の集合として識別結果文を得る．ここでの“良い文”とは，「文として自然である」ことを指す．類似候補文から，人手で良質な対訳コーパスを抽出することによって，翻訳性能が向上することは確認されているが[3]，性能の良い SMT の構築には大量の対訳コーパスが必要であり，識別の自動化は必須である．

識別器の素性として， N-gram[11]を用いる．N-gram DB には，出現頻度が保持されている．語句の置換えが発生した箇所を含む素性から，“良い文”“悪い文”の識別を行う「候補識別」器を構築する．「候補識別」器によって識別された文は，識別結果文として SMT の訓練文(対訳コーパス)となる．

識別器の具体的な処理を示す．まず，換言箇所への注視モデルを適用し，換言語句を最低 1 語含む N-gram フレーズをk個取得する．以下の式により，k個の各フレーズについて出現確率の対数尤度$Q(\omega_1\omega_2\cdots\omega_n)$を求める．

$$Q(\omega_1\omega_2\cdots\omega_n) \cong \log\prod_{i=1}^{n} P\left(\omega_i|\omega_{i-N+1}\cdots\omega_{i-1}\right)$$

k個の N-gram フレーズS_1~S_kの出現確率の対数尤度$Q_{s_t}(\omega_1\omega_2\cdots\omega_n)$の平均値$\overline{Q_{1k}}$によって文をスコア化し，閾値判定を行う．

Table 1: Corpora set

	観光/道案内ドメイン コーパス数(単位:文)
(1) 原文	13k
(2) 類似候補文	299k
(3) 識別結果文 (フィードバック:無し)	153k
(4-1) 識別結果文 (フィードバック:4,400文)	159k
(4-2) 識別結果文 (フィードバック:5,200文)	108k
(4-3) 識別結果文 (フィードバック:6,500文)	161k

※ それぞれの条件において、旅行コーパス450k文を含む

$$\overline{Q_{1k}} = \frac{1}{k}\sum_{t=1}^{k} Q_{s_t}(\omega_1\omega_2\cdots\omega_n)$$

平均値$\overline{Q_{1k}}$が閾値以上であれば，質の良い文と識別する．ただし，出現頻度から N-gram 確率を推定する場合，N-gram の学習モデル中に出現しない単語の確率値は 0 となってしまう．これを避けるために，予め N-gram の出現回数に一定の値を加えている．

$$P(\omega_n|\omega_{n-N+1}\cdots\omega_{n-1}) = \frac{C(\omega_{n-N+1}^{n}) + \delta}{C(\omega_{n-N+1}^{n-1}) + \delta V}$$

$C(\omega_1^n)$：単語列$\omega_1\omega_2\cdots\omega_n$が N-gram 中に出現する回数
V ：単語列の異なり総数
δ：定数($\delta = 0.5$)

3 フィードバック

本稿でのフィードバックとは，機械翻訳の利用者が翻訳を行った際に，得られた訳文に対して行う品質評価や，それに基づいて類似候補文や訳文の選択・修正を行う人的な作業(フォールバック)を意味する．

具体的には，SMT に入力文を与え，出力文(訳文)を得た際に，利用者が得られた訳文の品質を評価し，品質が低い場合に，良質な訳文と置換えを行うことや，入力文に近い類似候補文を提示し，その中から比較的品質の低いコーパス(意味的・文法的に破綻しているなど)を選択削除する等である(Fig. 3)．

本稿では特に，入力文に近い類似候補文を提示し，その中から比較的品質の低いコーパスをフォールバックとして選択削除する場合を取り上げる．この手法は，ユーザの原言語側でフィードバックを行うだけで済むため，コストを抑えながら，翻訳性能の向上を図ることができる．ここでは主に，識別器の N-gram DB の更新を行うことで識別性能を向上させることを目的としている．選

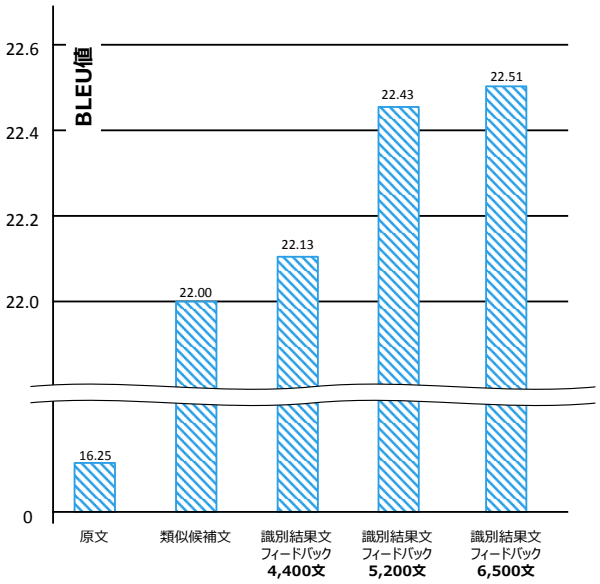

Fig. 4: BLEU score

択削除されずに残った比較的品質が高いと想定される類似候補文について，語句置換えが行われた語を最低1語含む N-gram フレーズk個に対し，それらの N-gram のスコアの算出に必要な各 N-gram 出現頻度に所定値を加算することで，更新を行う．N-gram DB に含まれない場合は，新たなフレーズとして N-gram DB に追加する．同様の処理で，選択削除された類似候補文から得られた N-gram フレーズに対しては，所定値を減算する．

これにより，ユーザのフィードバックを得るにつれ，識別器の N-gram DB のエントリ数が増え，口語表現等に対しても，正しい識別が可能になる．

また，フィードバックの質が悪い(“良い文”を，作業者が“悪い文”と誤判定する)場合を想定した際，不必要に対訳コーパスが削除されることを避けるため，選択削除された文の N-gram DB への更新寄与度は低く設定している．

4 実験・評価

フィードバックでの選択結果に基づいて，識別器を構成した場合の識別結果文から SMT を訓練する．その効果について，BLEU 値による客観評価，および主観評価を行う．

今回は，観光/道案内ドメインと仮定し，道案内における行動指示などで使われる言い回しを含んだ対訳コーパスで実験を行った．使用した対訳コーパスを Table 1 に示す．フィードバックの件数が少量(1～4,000 件)な場合においては，フィードバック数が増えるごとに翻訳性能が漸進的に向上することは確認されており[4], 本稿では，よりフィードバック件数を増やした場合の評価を行う．フィードバック件数は，それぞれ 4,400 文，5,200 文，

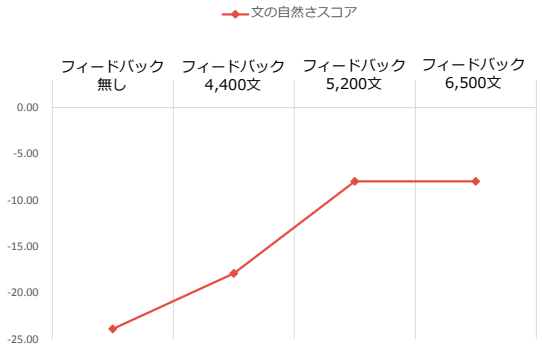

Fig. 5: Score of sentence accuracy

6,500 文であり，これらを元に識別器の N-gram DB を更新し類似候補文を識別後，翻訳モデルを再構築する．なお，フィードバック件数の増減と，識別候補文の増減は必ずしも一致しない．

なお，翻訳機の訓練では，識別結果文に加えて，汎用コーパスである旅行ドメインコーパス(約 45 万文対)を加えている．評価文は観光・道案内タスク文として約 75 文を用いた．当該の評価文は訓練文に含まれない．
今回は，識別器の素性として 4-gram を用いた．

4.1 客観評価

Fig. 4 に，各翻訳モデルの性能を示す(BLEU 値，10 回平均)．“類似候補文”に対して，“4,400 文”，“5,200 文”，“6,500 文”のそれぞれのフィードバック結果を比較すると，+0.13pt，＋0.43pt，＋0.51pt であり，“原文”と比較すると，+5.9pt，＋6.2pt，＋6.3pt となった．これにより，フィードバック件数が増加するに連れて，翻訳性能が向上する可能性が示唆された．特に，類似候補文と比較して，フィードバック後のコーパス数は 1/3～1/2 に減少しているのにも関わらず，翻訳性能が向上している．よって，翻訳性能を上げつつ，翻訳モデルのサイズを小さくできる可能性がある．

Fig. 5 にフィードバックの際に得られたスコアの一例を示している．対訳コーパス「牛乳は濃厚で大いに美味しいです.」(下線部が，とても → 大いに に置き換え)に対してのフィードバックごとのスコアの変化を示した．識別器はこのスコアが高いほど，“良い文”(文として自然)と判断する．この対訳コーパス自体は人間の目で見ると“良い文”であるが，N-gram DB のデータが不足しているためスコアが低い．しかし，対訳コーパスがフィードバックを得ると，直接フィードバックを受けていない場合(4,400 文)でもスコアが改善され，さらに，直接フィードバックを受けた際(5,200 文)には，大きなスコアの上昇が確認できた．

Table 2: Output examples

入力文1	**そのお店でオススメなのはシュークリーム！**
原文	At recommend the cream puff casier casier !
フィードバック (6,500文)	**At that store it's reccomend the cream puff!**
他の翻訳機	What you recommend at that shop is a cream puff!
入力文2	**おびひろ動物園は，北海道で２番目に誕生した動物園だよ**
原文のみ	The obihiro zoo was the second zoo to be born in Hokkaido it.
フィードバック (6,500文)	**The obihiro zoo was the second zoo to be born in Hokkaido.**
他の翻訳機	Hirohiro Zoo is the second largest zoo in Hokkaido
SMT(0)	**そこには山程おみやげがある**
原文のみ	There's something about the mountains there.
フィードバック (6,500文)	**There are many souvenirs.**
他の翻訳機	There are mountain souvenirs there.

4.2 主観評価

翻訳出力結果の事例を示す．入力文として，以下の各条件；

1. 想定ドメインで使われる言い回しを含む
2. 自然性の高い文(口語文調)
3. 原文・識別結果文に含まれない文

を満たす文として，3 文を挙げた．翻訳結果を Table 2 に示している．Table 2 では，原文をもとに構築した翻訳モデルによる出力結果を「原文」として示している．また，6,500 文のフィードバックを用いた識別器での翻訳モデルによる出力結果を「フィードバック(6,500 文)」と示している．加えて，一般的に利用可能な他の機械翻訳による翻訳結果を併記した．各訳出を比較すると，フィードバック適用後の翻訳例は，比較的良好な翻訳文の出力が確認できる．

5 さいごに

少量対訳コーパスからの統計的機械翻訳の構築を狙いとして，対訳コーパスを自動獲得する手法開発を行っている．類似候補文生成，候補識別を適用した統計的機械翻訳において，特に利用者からのフィードバックを活用し，識別器を順次更新することによる性能向上について報告し，特に BLEU 値(最大値)で，＋6.3 ポイントの向上効果を得た．

謝辞

本研究の一部(汎用コーパスの活用)は総務省の情報通信技術の研究開発「グローバルコミュニケーション計画の推進-多言語音声翻訳技術の研究開発及び社会実証- I. 多言語音声翻訳技術の研究開発」の一環として行われました．

参考文献

[1] 松田他: “多言語音声翻訳システム”VoiceTra”の構築と実運用による大規模実証実験”， 信学 D， No.10，pp.2549-2561(2013)

[2] KOEHN P.， “Statistical Phrase-Based Translation: Proc. Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics”，HLT-NAACL-03， (2003)

[3] 藤原他， “自動コーパス生成による少量対訳コーパスからの統計的機械翻訳”， 言語処理学会第 22 回大会 (2016)

[4] 山内他， “自動コーパス生成とフィードバックによる少量対訳コーパスからの統計機械翻訳”，2016 年度人工知能学会 全国大会(2016)

[5] 藤原他， “コーパスの自動生成･識別による少量コーパスからの統計的機械翻訳”， 第 15 回情報科学技術フォーラム(2016)

[6] Madnani N.et al， “Generating targeted paraphrases for improved translation”， ACM Trans. Intell. Syst. Technol.4， 3， Article 40 (2013)

[7] Yuval M，et al.， “Distributional Phrasal Paraphrase Generation for Statistical Machine Translation”， ACM Trans. Intell. Syst. Technol.4， 3， Article 39 (2013)

[8] Japanese Wordnet (v1.1)，http://compling.hss.ntu.edu.sg/wnja/

[9] Mizukami M et al.， “Building a Free，General-Domain Paraphrase Database for Japanese”，The 17th Oriental COCOSDA Conference (2014)

[10] 山形他， “普通名詞換言辞書の構築”， 言語処理学会第 20 回年次大会， pp.7-10 (2014)

[11] http://s-yata.jp/corpus/nwc2010/ngrams/