2P1-8in1

自動コーパス生成とフィードバックによる少量コーパスからの統計機械翻訳

Statistical Machine Translation using Small Parallel Corpora based on Automatic Corpora Generation and Feedback

山内 真樹[*1,*2] Masaki Yamauchi　藤原 菜々美[*1] Nanami Fujiwara　内山 将夫[*2] Masao Uchiyama　隅田 英一郎[*2] Eiichiro Sumita

[*1] パナソニック株式会社 先端研究本部 Advanced Research Division, Panasonic Corporation　[*2] 情報通信研究機構 NICT

Recently Statistical machine translation (SMT) systems have been widely used. SMT is a system whose model is trained on large quantities of parallel corpora. Although the size of the corpora directly affects performance of SMT, corpora data is expensive in general. From this perspective, we have been developing a unique method which automatically generates large-size parallel corpora from a small number. The method consists of “creating candidate corpora by utilizing various expressions and paraphrases knowledge” and “choosing better corpora by a discriminator”. In this letter, we report that SMT performance is improved more than three points compared to SMT with original small corpora, by using translation feedbacks to generate supervised data as positive or negative score for the candidate corpora which are used for inputting to the discriminator.

1. はじめに

大量の対訳コーパスから，翻訳に必要なモデルを統計的に獲得する統計的機械翻訳システム(SMT: Statistical Machine Translation) [KOEHN 03] が登場している．欧州言語間など言語・文法構造が近い言語間では，SMT による機械翻訳が実用域に達しつつある．日本語を中心とした翻訳(日英間，日・アジア言語間等)でも，利用ドメインを「旅行会話」などに限定することにより実証実験段階となっている領域がある[松田 13]．

一方，新規ドメイン向けに翻訳機を構築する場合は，新たに対訳コーパスが必要となる．SMT の構築には大量の対訳コーパスが必要であるが，新規ドメインでの大量コーパスの収集は一般に困難であり，特に初期段階で準備できる対訳コーパス量は，ドメインに依らずおおよそ 1,000～10,000 文オーダ前後となる．少量の対訳コーパスでは統計的に十分な情報が得られず，SMT の性能は著しく低下するため，このような状況下での翻訳エンジン構築は極めて挑戦的な課題である．

これに対し我々は，少量の対訳コーパスからの統計的機械翻訳(翻訳エンジン)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[藤原 16]．翻訳性能を向上しつつコーパス生成のコスト削減を図るため，種となる少数の対訳コーパスから類似候補文を生成し，機械学習により好ましい文を識別することを狙いとしている．本稿では，類似候補文の自動生成と統計的機械翻訳への適用による翻訳性能に関して，特に利用者からのフィードバックを活用し，好ましい文を選択する識別器を順次更新することによる漸進的な性能向上手法の検討とその評価について報告する．

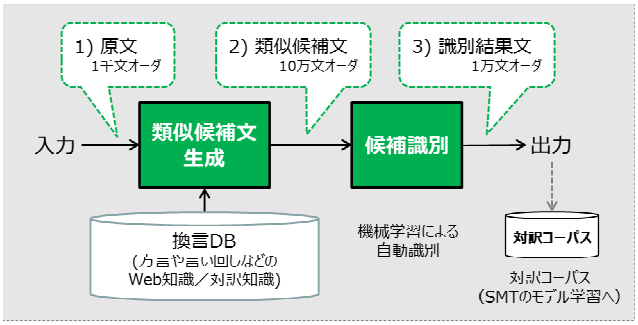


Fig. 1 Automatic Corpora Generation

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

2. 自動コーパス生成：ACG

本項では自動対訳コーパス生成手法について説明する．SMT は，対訳コーパスを基に統計的に得られた確率分布から推定を行うため，翻訳性能がコーパスの質・量に依存する．コーパス追加等での性能評価の際は確率分布の変化に留意する．

我々が開発している ACG の構成概要図を Fig. 1 に示す．ACG は，対訳コーパス入力から「類似候補文生成」器と「候補識別」器により多量のコーパス(識別結果文)を生成する．

「類似候補文生成」器では，言換え表現のデータベースを言語資源(WordNet [Word 09], PPDB[Mizukami 14] ,内容語換言辞書[山形 14]等)，及び手作業から構築し(換言データベース)，入力文に対して類似候補文を生成する．

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

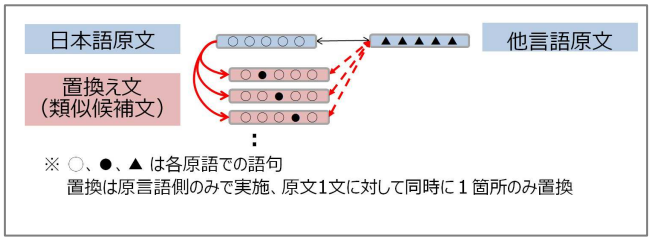


Fig. 2 Original and candidate corpora

コーパスに適切な文を識別する．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．

類する先行研究としては，WordNet から言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[Madnani 13]や，置換えルールでのコーパス拡張手法[Yuval 13]などが挙げられる．

本稿では「類似候補文生成」器による類似候補文に対して，選択的に識別器を適用・学習することによる翻訳機の漸進的な性能向上を目的として，翻訳結果に対するフィードバックを基に，識別対象とする候補文の抽出と識別についての仮説と，「候補識別」器による識別の際に期待される効果について評価を行う．

3. 候補識別とフィードバック

3.1 識別

類似候補文に対して識別器を適用することで，”良い文”の集合として識別結果文を得る．次段で述べるフィードバックにより類似候補文の識別を可変とし，且つ，識別結果文により構築する翻訳モデルの性能評価を容易なものとするため，ここでは識別器の構成を単純化する．識別用の素性を n-gram の出現頻度[矢田 10]とし，語句置換えが発生した箇所を含む n-gram 出現頻度をスコアとした閾値判定を行う．具体的には，各類似候補文について採用もしくは棄却の 2 値判定とし，識別閾値の初期条件として，類似候補文中の n-gram のうち，スコアが所定値以下の配列が規定個数以上発生した場合に，その類似候補文を棄却する．

3.2 フィードバック

本稿でのフィードバックとは，SMT 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，利用者がその訳文の品質を評価し，SMT 出力の信頼度が低いとされた場合に他の方法[*1] で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)を選択削除することを指す．概要図を Fig. 3 に示す．

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．選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，その n-gram スコアを各々求め，スコアが所定値以下の配列が存在した場合，その配列に対するスコアを所定値より大きい値と置き換える．これにより，同じ n-gram を有する類似候補文に対して，品質が良い類似候補文として識別される確率を上げることを目的とする．

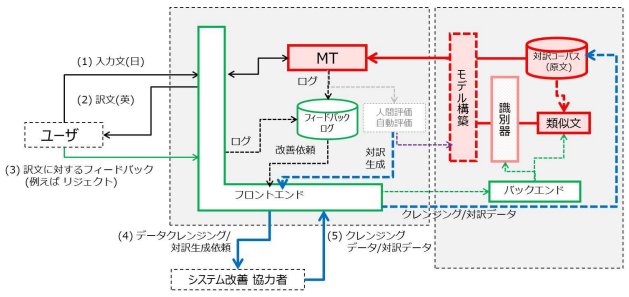


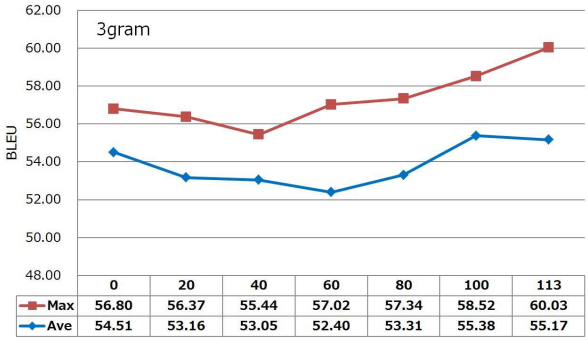
Fig. 3: Feedback system

＊ 1: 入力文に対する訳文を他の利用者や協力者による翻訳結果から得ること，機械翻訳結果に対して人間による選択・修正を得ることなど

Table 1: Corpora Set

	学習コーパス
(1) 原文	0.1K
(2) 類似候補文	8.1K
(3) 識別結果文	可変(0.1～4.4k)

※ 翻訳モデル訓練時に道案内コーパス5.0Kと旅行コーパス0.16Mを含む



	0	20	40	60	80	100	113
Max	56.80	56.37	55.44	57.02	57.34	58.52	60.03
Ave	54.51	53.16	53.05	52.40	53.31	55.38	55.17

Fig. 4: Feedback sentences and BLEU score for 3-gram

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．その効果を BLEU 値にて評価することで効果を定量的に確認する．

具体的な少量コーパス例として，対象ドメインを道案内タスクと仮定し実験を行った．使用したコーパスを Table 1 に示す．少量コーパスとして，原文に道案内における行動指示で使われる言い回しを含んだ対訳コーパス(約 100 文対，日英対訳文)を用いた．この原文に対し ACG[藤原 16]を適用し，類似候補文を得た．類似候補文の文数は約 8,100 文である．

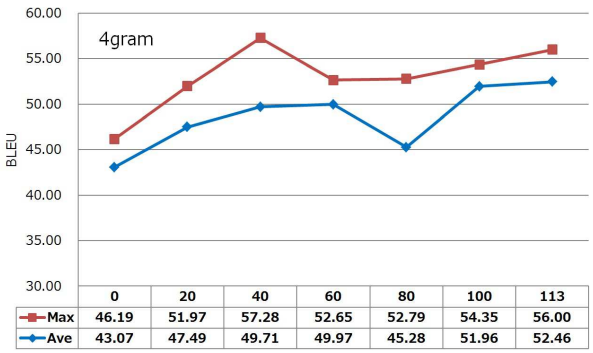
SMT の翻訳モデル訓練に用いる識別結果文は，識別器の状況，すなわちフィードバックが発生した状況に依り変化する．今回の実験では，識別結果文の文数として概ね 100 から 4,400 文を得た．なお，翻訳機の訓練では，識別結果文に加えてベース用コーパスとして，旅行ドメインのコーパス(約 16 万文対)，及び，道案内コーパス(約 5,000 文対)を加えている．評価文は道案内タスクから 30 文を抽出して用いた．当該の評価文は訓練文から削除して訓練を行っている．

また，識別器における n-gram スコアの所定値は 0 とし，規定個数は 2 として実験を行った．

Fig. 4, Fig. 5, Fig. 6 に，識別器の n-gram 素性としてそれぞれ 3-gram, 4-gram, 5-gram を用いた場合における，識別結果文を用いての翻訳モデル性能を示す (BLEU 値)．

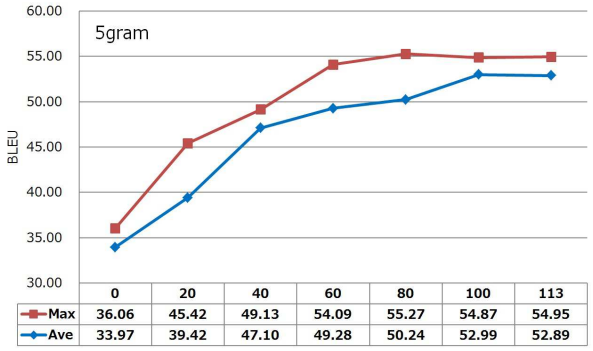
横軸はそれぞれフィードバック文数を表している[*2]．各翻訳モデルの訓練ではチューニングによる影響を考慮し，各条件において 10 回ずつのモデル生成を行っている．それぞれのグラフ中で，ave は 10 回の平均 BLEU 値を，max は 10 回中の最高 BLEU 値を表している．フィードバックの最大数は，全ての例において 113 文である．

＊ 2: 例えば ”20” は，「入力文 20 文に対してフィードバックを行い，入力文に近い類似候補文をそれぞれ提示，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した場合(選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，その n-gram スコアを所定値より大きい値と置き換えた場合)」に，「フィードバック後の n-gram スコアを用いて識別器を構築」し，「その識別器を用いて類似候補文の識別を行った結果として得られた識別結果文」を用いて翻訳モデルを訓練した場合を表している．



	0	20	40	60	80	100	113
Max	46.19	51.97	57.28	52.65	52.79	54.35	56.00
Ave	43.07	47.49	49.71	49.97	45.28	51.96	52.46

Fig. 5: Feedback sentences and BLEU score for 4-gram



	0	20	40	60	80	100	113
Max	36.06	45.42	49.13	54.09	55.27	54.87	54.95
Ave	33.97	39.42	47.10	49.28	50.24	52.99	52.89

Fig. 6: Feedback sentences and BLEU score for 5-gram

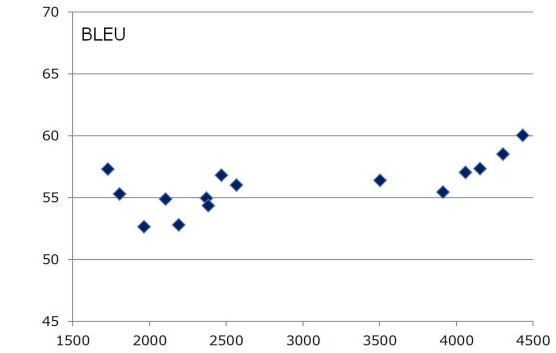


Fig. 7: BLEU vs. Number of "selected" corpora by discriminators using feedbacks

4.1 客観評価

得られた BLEU 値について考察する. 原文に対してフィードバック無しの場合(0 文)と, 全原文に対してフィードバックした場合(113 文)とを比較する. Fig. 4, Fig. 5, Fig. 6 の 3-gram, 4-gram, 5-gram において, 平均(ave)で 0.66~18.92 ポイント, 最高値(max)で 3.23~18.89 ポイントの BLEU 値改善が得られた. 平均値・最高値ともに良化が確認されるとともに, 最高値においては 3-gram, 4-gram, 5-gram のいずれの場合においても大きな改善が確認でき, 特に 5-gram の場合において顕著な効果が見られる. 但し, 3-gram, 4-gram の場合と比較して, 5-gram の場合は初期条件(0 文)での BLEU 値が低いため, フィードバック効果が相対的に強く出ている可能性が有る.

Table 2: Output examples

入力文1	**右側にまっすぐ進みなさい。**
SMT(0)	And go straight down on the right side.
SMT(113)	**Go straight on the right side.**
他の翻訳機例	Please go straight to the right.
入力文2	**左斜め前にエスカレーターがございます。**
SMT(0)	There is an escalator diagonally on your left side ahead of you.
SMT(113)	**There is an escalator diagonally left before.**
他の翻訳機例	There is a escalator before left oblique.
SMT(0)	**進んだ先にコンビニがあります。**
SMT(0)	First you will see a convenience store at the end.
SMT(113)	**After you continue there is a convenience store.**
他の翻訳機例	There is a convenience store in the advanced ahead.
入力文4	**正面の出口から外へ出て、信号を渡ってください。**
SMT(0)	Please cross at the traffic light from from the exit in front of you.
SMT(113)	**Go out from the exit in front of you please cross at the traffic light.**
他の翻訳機例	Out from the outlet of the front to the outside, please across the signal.

全原文に対してフィードバックを行った場合(113 文)で比較すると, 3-gram が 4-gram, 5-gram と比して良化傾向にある. 他方, 平均値と最高値の差からも推察できるように, チューニングの揺らぎによる影響も少なくないことが推測される. 全体としては, 識別器素性に言語モデルを用いた場合において, フィードバックによる改善効果があると考えられる.

なお, 識別結果文の総数は, フィードバック対象となる原文数が増えるに伴って増加傾向にある. このため, 翻訳モデルの訓練時における対訳コーパス総数の増加に伴って BLEU 値が向上している可能性がある.

Fig. 7 に,「フィードバックで構築・更新した識別器に依って選択された識別結果文の総数」と BLEU 値との関係を示す. 今回実験を行った範囲におけるフィードバック・識別器構築では, 対訳コーパス数と BLEU 値の間に強い関連は認められず, 今回の BLEU 値の改善は, フィードバックにより識別器が効率良く識別結果文を生成できていることによる可能性も示唆される.

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, 4 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 SMT(0)として示している. また, 113 文に対する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 SMT(113)として示している. また一般的に利用可能な他の機械翻訳機による翻訳結果を併記している. SMT(0)と SMT(113)の訳文を比較すると, フィードバックに基づいた識別結果文による SMT(113)では, 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推定・獲得する手法開発を行っている.

類似候補文の自動生成, フィードバックによる識別器構築での識別結果文生成, 及び翻訳性能の向上について報告し, 特に BLEU 値(最大値)で, 約 3.23~18.89 ポイントの向上効果を得た.

謝辞

本研究の一部(フォールバックでの人間による修正)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

[KOEHN 03] 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)

[松田 13] 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)

[Papineni 02] Papineni K. et al, BLEU: a method for automatic evaluation of machine translation: 40th Annual meeting of the Association for Computational Linguistics p.311-318 (2002).

[Madnani 13] Madnani N.et al, Generating targeted paraphrases for improved translation. ACM Trans. Intell. Syst. Technol.4, 3, Article 40 (2013)

[Yuval 13] Yuval M,et al., Distributional Phrasal Paraphrase Generation for Statistical Machine Translation. ACM Trans. Intell. Syst. Technol.4, 3, Article 39 (2013)

[Word 09] Japanese Wordnet (v1.1): http://compling.hss.ntu.edu.sg/wnja/

[Mizukami 14] Mizukami M et al., Building a Free, General-Domain Paraphrase Database for Japanese: The 17th Oriental COCOSDA Conference (2014)

[山形 14] 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)

[藤原 16] 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)

[矢田 10] http://s-yata.jp/corpus/nwc2010/ngrams/