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

藤原 菜々美† 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 知能研究室
‡情報通信研究機構 先進的翻訳技術研究室

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

1 はじめに

近年，大量の翻訳データ(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)[1,2]が登場している．SMTは，大量の対訳コーパス(原言語と目的言語の文章対データ)から，翻訳に必要なモデルを統計的に獲得する．

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，利用領域(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[3]がある．

一方で，翻訳言語対を一つ追加(対応言語を一つ増やす)するごとに，先の旅行用翻訳であれば百万文オーダの翻訳文が必要であり，費用は数億円/一言語対にも上る．また，新規の利用領域(ドメイン；飲食店・小売・交通・宿泊施設等)向けに翻訳機を構築する場合には，翻訳文だけではなく，そのドメイン特有の表現を含んだ原文が必要となり，更にコストは倍増する．

しかし，新規のドメインで大量の原文・対訳コーパスを収集することは一般に困難である．特に初期段階では，準備できる対訳コーパス量は，各ドメインおおよそ1,000文オーダ前後となる．1,000文オーダといった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の性能も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，機械翻訳において極めて挑戦的な課題である．更に，対訳コーパス数の衆寡が翻訳性能に直結する一方で，対訳コーパスの収集・獲得は高コストであり，SMTの実用化課題でもある．

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．翻訳性能を向上しつつ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(識別学習)により類似候補文から正しい類似コーパスを自動で獲得することを狙いとしている．

本稿では第一報として，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の向上について報告する．

2 システム構成

本項では統計的機械翻訳(SMT)システムの全体構成と，新規に開発を行っている自動対訳コーパス生成手法について説明する．

2.1 統計的機械翻訳：SMT

我々が用いているSMTの構成概略図をFig.1に示す．簡単のため，二つの構成要素に分け説明する．ひとつは事前に統計的な翻訳モデル・言語モデルを構築する「モデル学習」，もうひとつは事前に構築されたモデルに従い，最尤推定により入力文(原言語)から確率的に最適と推定される訳文を出力文(目的言語)として出力する「翻訳エンジン」である[2]．

「モデル学習」では，対訳コーパス・単言語データを用いて統計的に翻訳モデル・言語モデルを構築する．原言語文をJ，目的言語文をEとすると，原言語から目的言語への翻訳は確率$P(E|J)$の最大化タスクとなり，ベイズの定理から次式となる；

$$P(E|J) = \frac{P(J|E)P(E)}{P(J)} \propto P(J|E)P(E)$$

言語モデルは$P(E)$に相当する．目的言語らしさ(流暢さ)を表す確率と考えられ，単言語(目的言語)の文データから統計的に獲得する．翻訳モデルは$P(J|E)$に相当する．“ある目的言語文(単語/句)が，ある原言語文(単語/句)であった確率”と考えられ，対訳コーパスから統計的に獲得する．

「翻訳エンジン」では，翻訳モデル及び言語モデルをもとに，目的文候補を最尤復号する．統計的に得られた確率分布をもとに推定を行うため，SMT性能はコーパスの質・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．

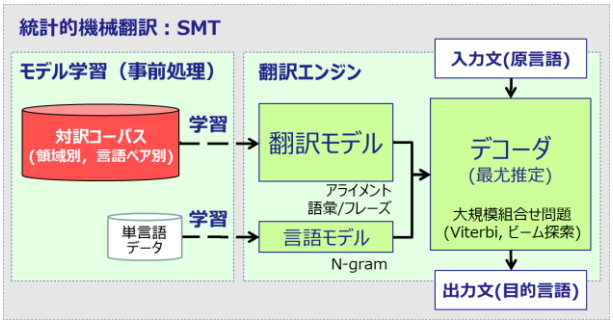

Fig. 1 SMT system

2.2 自動対訳コーパス生成：ACG

我々が開発している ACG の構成概要図を Fig.2 に示す. ACG は，少量の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)を自動生成する．「類似候補文生成」器では，表現形式・言い回しを言い換え表現のデータベース(換言 DB)として Web・各種言語辞書(WordNet[7],PPDB[8] ,内容語換言辞書[9]等)・手作業でのデータ構築等により構築する．言い換え元となるフレーズと，そのフレーズを置き換える別のフレーズ群とを対応付けた辞書として構成しており，入力文の中に DB エントリと一致するフレーズが有った場合に，そのフレーズを置き換えて類似候補文として生成する．

換言 DB を用いて生成した類似候補文は，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 DB 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，このような破綻した類似候補文を機械学習により除外し，対訳コーパスとして適切な文を自動的に識別し，識別結果文を得る．

類する先行研究としては，Madnani ら[5]による 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や，Yuval ら[6]による置換えルールでのコーパス拡張手法などが挙げられる．一方，機械翻訳の学習データとなる対訳コーパスについて，類似候補文を生成し機械学習により自動で識別を行う枠組みは，近い事例が少なく報告例を見いだせていない．

我々は現在，ニューラル・ネットワークを用いた候補識別器を設計中である．この詳細については次報とし，本稿では第一報として「類似候補文生成」器による影響と，「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．

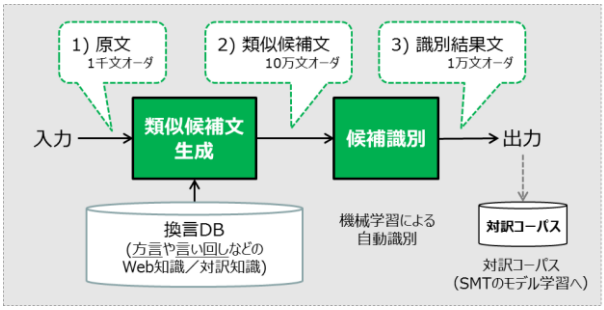

Fig. 2 Automatic Corpora Generation

3 評価

少量対訳コーパスを用いた SMT 構築について，自動対訳コーパス生成の類似候補文及び，識別結果文が与える影響・効果を，SMT 出力文(訳文)から評価を行う．評価は，以下の各コーパスから SMT を学習・構成した場合の翻訳性能をもとに行う．

1）少量の原対訳コーパス(原文)

Table 1: Factory and Medical Corpora Sets

	工場コーパス	医院コーパス
(1)原文	**0.5K**	**0.23M**
(2)類似候補文※	**4.8K** + (1)原文	**6.5K** + (1)原文
(3)識別結果文	**0.8K** + (1)原文	**2.0K** + (1)原文

※ 工場コーパス，医院コーパスにおいて，それぞれ20文，200文を抽出し類似候補文を生成した．ベースとして全てに旅行コーパス0.16Mを含んでいる．

2）本方式(ACG)により生成した対訳コーパス(類似候補文)

3）類似候補文を人手でクレンジングした対訳コーパス
(識別結果文に相当，本来は識別器による識別を想定)

具体的なコーパスとして，対象ドメインを工場及び医院診察と仮定し，原文に工場及び医院診察で使われる言い回しを含んだ対訳コーパス(工場：約500文対，日英対訳文、医院：約23万文対，日英対訳文)を用意した．それぞれの条件において，工場及び医院コーパスに，旅行コーパス約16万文対をベースとして加えている．各コーパスの文数は Table1に示している．

工場コーパスについては，評価用として原文から約20文をランダムに抽出し，評価用コーパスとした．SMT の学習には，原文・類似候補文・識別結果文を用い，評価用コーパスは除いている．医院コーパスについては，評価用として原文からランダムに200文を抽出し，表現を一部変えたものを評価用コーパスとした．工場コーパスと同様に、SMT の学習として，原文・類似候補文・識別結果文を用い，評価用コーパスは除いている．

なお，各コーパスの SMT の学習において，各コーパス(工場・医院)の類似候補文と識別結果文の文数を同程度に合わせるため，SMT の学習時には識別結果文は同じ文を重複して含んでいる．また，識別による効果を推定するため，類似候補文に含まれる“破綻した文”について，今回は人手でクレンジングを行い，これを識別結果文とした．ここでのクレンジングとは，文法的に正しく，原文と意味が同じ文の抽出である．

3.1 翻訳結果評価方法

3.1.1 新規コーパス生成効率

SMT 性能は BLEU[4]での評価が一般的である．一方，評価文が互いに異なる場合, SMT 性能を単純に BLEU 値で比較できない．先行研究との比較のために，本稿の主旨に沿い，「どれだけの対訳コーパスを新規に生成できたか」を評価軸とする．コーパス数と BLEU 値には相関があり，BLEU 値を対訳コーパス数に換算できる[10]. 換算された対訳コーパス数と原文数との比率を新規コーパス生成効率 p として，次式より算出する※2.

$$y = 40.584\, e^{0.123x}$$

(x：BLEU 値，y：換算コーパス数．100文~10万文の対訳コーパスによる BLEU スコア分布から近似的に導出．$R^2 = 0.9676$)

$$p = \frac{(N + y_2 - y_1)}{N}$$

(N：原文数，y_1，y_2：換算コーパス数，p：新規コーパス生成効率)

Table 2: SMT and ACG Performances of the Factory Corpora Sets

工場コーパス		BLEU値*	コーパス生成効率（倍）	
(F-1)	少量の原対訳コーパス（原文）	7.21		
(F-2)	本方式(ACG)により生成した対訳コーパス（類似候補文）	20.86	(2)-(1) 8.33	
(F-3)	類似候補文をクレンジングした対訳コーパス（識別結果文）	22.03	(3)-(1) 10.53	(3)-(2) **3.20**
医院コーパス		BLEU値*	コーパス生成効率（倍）	
(M-1)	少量の原対訳コーパス（原文）	11.60		
(M-2)	本方式(ACG)により生成した対訳コーパス（類似候補文）	21.55	(2)-(1) 11.44	
(M-3)	類似候補文をクレンジングした対訳コーパス（識別結果文）	22.74	(3)-(1) 14.32	(3)-(2) **3.89**
先行技術		BLEU値*	コーパス生成効率（倍）	
(4)	先行技術[7] スペイン語と英語の対訳対	ー	1.16	
(5)	先行技術[8] 英語と中国語の対訳対	ー	1.45	

3.2 自動対訳コーパス生成 (ACG) 評価

評価結果を述べる．本評価では，コーパス生成の翻訳性能への寄与を評価するため，以下の2つの観点から評価を行う．

- 客観評価：BLEU 値をもとにコーパス生成効率より評価
- 主観評価：翻訳出力結果事例をもとに評価

3.2.1 客観評価：コーパス生成効率

対訳コーパス質・量の観点から評価する(Table2)．原文・類似候補文・識別結果文の各々における翻訳性能をBLEU値で，ACGによるコーパス生成効率を増減比で示している．

工場コーパスにおいて，原文と類似候補文との比較では，BLEU 値で+13.65，コーパス生成効率で約8.33倍となり，類似候補文と識別結果文との比較では，BLEU 値で＋1.17，コーパス生成効率として約3.20倍、原文と識別結果文では，＋14.78，約10.53倍となった．また，医院コーパスにおいて，原文と類似候補文との比較では，BLEU 値で+9.95，コーパス生成効率で約11.44倍となり，類似候補文と識別結果文との比較では，BLEU 値で＋1.19，コーパス生成効率として約3.89倍、原文と識別結果文では，＋11.14，約14.32倍となった．

[先行例] Madnani ら[5]による手法では，コーパス生成効率は約1.16倍であった(英・スペイン語)．Yuval ら[6]による手法では，コーパス生成効率は約1.45倍であった(中・英語)．

[考察] 上記から，ACG による類似候補文によって翻訳性能が向上することが示された．類似候補文と識別結果文との比較からは，破綻した文を削除し正しい文を選択的に識別することが，翻訳性能を向上させることを強く示唆している．

翻訳性能に寄与する類似文の生成効率という観点では，先行手法と比較して，我々の ACG による類似候補文生成は効率が良いと言える．先行例では有意なフレーズ置換えのみを事前に選択した上で適用しているが，ACG では品質を問わず置換えを実行している．このため，SMT に有効なコーパスを多く獲得できたものと推測される．

コーパス識別の効果という観点では，ACG で品質を問わず生成した類似候補文を事後的に識別することで，更に生成効率を高められることが示唆されている．事前に適用ルールを判断する先行例では，システム全体での効率が1.16～1.45倍程度であるのに対し，生成後に識別を行う ACG では識別効果として約3.20～3.89倍が今回の実験結果では示された．

さらに本アプローチは，適用するドメインに関わらず，汎用的に性能を向上できることが示された．また，少量の対訳コーパスだけでなく，比較的多くの文を有する場合でも，有効である可能性が示唆された．

これらより，本手法のアプローチが有意であると言える．

Table 3: Translation Examples

入力文1	**この製品を今後、ヨーロッパに販売していきたいんだけど。**
(1) での翻訳	I'd like to buy the products schedule europe'll have this.
(3) での翻訳	**I'd like to sell this product in europe little schedule.**
他の翻訳機	The future of this product, and I want to continue to sell to Europe.
入力文2	**価格競争力の向上も大事だ。**
(1) での翻訳	I improve the priceis competitive care.
(3) での翻訳	**Its is almost important to improve price competitive.**
他の翻訳機	Also a important improvement in the price competitiveness.
入力文3	**これは高めなんかなぁ。**
(1) での翻訳	This is a little expensive for example.
(3) での翻訳	**This is high about it.**
他の翻訳機	This wonder if do a raise.
入力文4	**体調がおかしくなったら言ってくれ。**
(1) での翻訳	When you say haywire well?
(3) での翻訳	**Tell me if you feel hard well.**
他の翻訳機	Let me say Once the physical condition becomes amusing.
入力文5	**会社の健康診断で悪い結果が出てしまったよ。**
(1) での翻訳	The result of medical checkup in the company you got sick.
(3) での翻訳	**I've got a bad result company's medical checkup.**
他の翻訳機	Bad results had come out in the health diagnosis of the company.
入力文6	**私は大丈夫なんやろか。**
(1) での翻訳	I have an okay.
(3) での翻訳	**Is it okay to me?**
他の翻訳機	Do Yaro do I okay.
入力文7	**外出してもいいん？**
(1) での翻訳	Can I'm out?
(3) での翻訳	**Can I go out?**
他の翻訳機	It does good to go out ?
入力文8	**ここでお薬は出してもらえるんですか？**
(1) での翻訳	Here is your medicine is fast.
(3) での翻訳	**Do you have a medicine here?**
他の翻訳機	Where do I get to your medication issues ?
入力文9	**そん通り飲んでみます。**
(1) での翻訳	I'll try that street.
(3) での翻訳	**I'll try that right.**
他の翻訳機	I'll try drinking Son Street
入力文10	**血圧はが少し高いかもしれへん。**
(1) での翻訳	My blood pressures was a little high strange.
(3) での翻訳	**It's alittle high blood pressure.**
他の翻訳機	Hen may have high blood pressure is a little.

※2 換算対訳コーパス数のみの比較では，新規に生成されるコーパス数が原文数に依存する可能性がある．原文数との比率で生成効率を算出することで原文数に依らない比較を可能とする．生成された対訳コーパス数に対して種コーパスが少量であるほど，対訳コーパスが効率良く生成できており．コーパス生成効率は高いほど良いと言える．

3.2.2 主観評価：翻訳出力結果の事例

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(工場，及び医院診療)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，10文を挙げた．各翻訳エンジンによる翻訳結果を Table 3 に示している．

[考察] 原文をもとに構築した SMT: (1) や，Web 上での一般的に利用できる翻訳機での翻訳結果を，識別結果文をもとに構築した SMT: (3) での翻訳結果と比較すると，(3) の識別結果文をもとにした SMT では，比較的良好な翻訳文を出力することができている．

例えば，入力文5「会社の健康診断で悪い結果が出てしまったよ。」のようなドメインに特化した表現が含まれる文に対して，Web 上の翻訳機では「会社の健康診断」という語を正しく翻訳することができていない．SMT: (1)では，ドメインに特化した表現は翻訳できているものの，全体として正しい訳を出力できなかった．一方，SMT: (3)ではドメイン特化の表現を訳出した上で，全体として正しい翻訳を出力することができている．入力文6「私は大丈夫なんやろか。」のような表現に対しても，Web 上の翻訳機では「やろ」が「Yaro」と翻訳されたのに対し，このような表現を学習した SMT: (3)においては，正しい翻訳が得られている．入力文9「そん通り飲んでみます。（”その通り”の口語表現）」のような表現が含まれた文に対しても，SMT: (1)や，Web 上の翻訳機では「通り」を「street」とする誤訳となっているが，口語的な表現を学習した SMT: (3)では，そのような意味の誤訳をすることなく，比較的良好な翻訳文を出力することを可能としている．

識別結果文では，口語的な表現，ドメインに特化した表現からモデルを学習していることが寄与し，意味の通じる翻訳文を出力することができたと考えられる．

4 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で推定・獲得する手法開発を行っている．

類似候補文の自動生成による翻訳性能の向上について報告し，特に，翻訳性能を対コーパス数比で換算した場合に，事前に適用ルールを判断する先行例では1.16～1.45倍程度であるのに対し，生成後に識別を行う本手法では暫定値として約3.20～14.32倍以上と顕著な効果を得た．

今後は，工場・医院診療以外のドメイン(例えば，小売や鉄道など)でも評価を行い，その有用性と類似候補文を含むことによる悪影響の有無を検証する．さらに，今回は人手で行った類似候補文のクレンジングに関しても，候補識別器を構築する予定である．

5 参考文献

1) 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)
2) Moses, statistical machine translation system http://www.statmt.org/moses/
3) 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)
4) 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).
5) Madnani N.et al, Generating targeted paraphrases for improved translation. ACM Trans. Intell. Syst. Technol.4, 3, Article 40 (2013)
6) Yuval M,et al., Distributional Phrasal Paraphrase Generation for Statistical Machine Translation. ACM Trans. Intell. Syst. Technol.4, 3, Article 39 (2013)
7) Japanese Wordnet: http://compling.hss.ntu.edu.sg/wnja/
8) Masahiro M et al., Building a Free, General-Domain Paraphrase Database for Japanese: The 17th Oriental COCOSDA Conference (2014)
9) 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)
10) 岡田他: 文分解法によるコーパス規模と Bleu 値との関係: 平成 20 年度 AAMT/Japio 特許翻訳研究会 報告書 pp.14-22(2008)