E-014

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

藤原 菜々美† 山内 真樹†
Nanami FUJIWARA Masaki YAMAUCHI

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) [1] 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[2].

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[3][4]．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，類似候補文中から良質な対訳を自動で識別(選択)した時の翻訳性能について報告する．

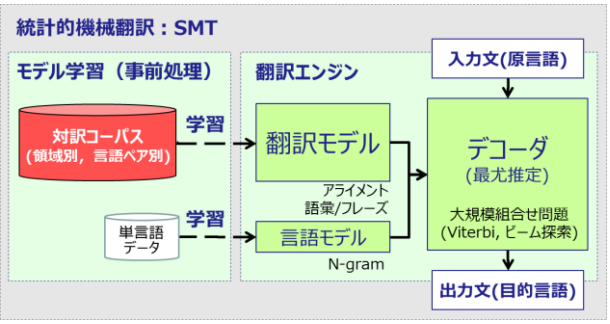


Fig. 1 SMT system

†パナソニック株式会社 先端研究本部

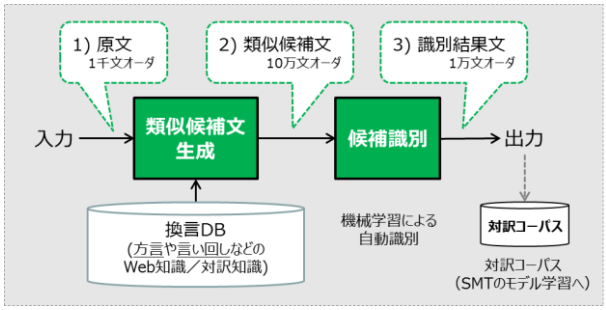


Fig. 2 Automatic corpora generation

2. システム構成

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[5].

「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．

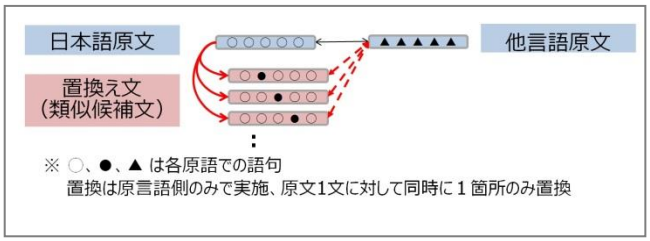


Fig. 3 Original and candidate corpora

「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．

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，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．

2.2.1 類似候補文生成

「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 [6], PPDB[7] ,内容語換言辞書[8]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．

類似候補文の生成模式図を Fig. 3 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．

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[9]や，置換えルールでのコーパス拡張手法[10]などが挙げられる．

2.2.2 候補識別

「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．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，識別の自動化は必要不可欠である．

識別器の素性として， N-gram[11]を用いる．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．

Table 1: Evaluation corpus sets

換言DB サイズ	ⅰ)1.0Mエントリ	ⅱ)2.5Mエントリ
(1) 原文	-	-
(2) 類似候補文	130K	410K
(3-1) 識別結果文 (正規化無)	60K	220K
(3-2) 識別結果文 (正規化有)	125K	440K

＊ それぞれの条件において、旅行コーパス0.45M，原文5.0Kを含んでいる

識別器の具体的な処理として，ある類似候補文 1 文において，語句置換えが行われた語を最低 1 語含む N-gram を k 個取得する．k 個のそれぞれについて N-gram の出現確率の対数尤度を求め、平均値$\bar{k}$に対して閾値判定を行う．語句置換えが行われた箇所の周辺のフレーズに対し，それらがある一定以上の出現確率を持つ場合に，“良い文”として識別結果文となる．

N-gram の出現確率は，単語$\omega_1\omega_2\cdots\omega_n$の出現確率を$P(\omega_1\omega_2\cdots\omega_n)$として，以下の式で求める．

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

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も多く存在するため，学習モデルに含まれない単語の確率値が 0 となる場合がある．そのため，加算スムージングを行い，出現確率を求める際に，N-gram の出現回数に一定の値を加えている．

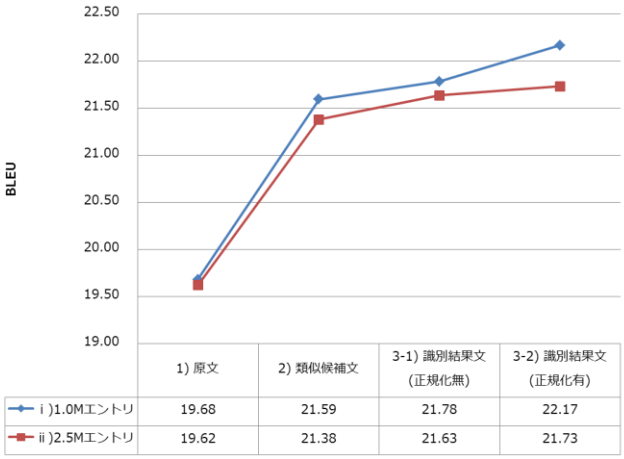
$$P(\omega_i|\omega_{i-N+1}\cdots\omega_{i-1}) = \frac{C(\omega_{n-N+1}^{n}) + \delta}{C(\omega_{n-N+1}^{n-1}) + \delta V}$$

ここで，$C(\omega_1^n)$は単語列$\omega_1\omega_2\cdots\omega_n$が N-gram の学習データ中に出現する回数，$V$は単語列の異なり総数，$\delta$は定数$(\delta = 0.5)$である．

本稿では「類似候補文生成」器により生成された類似候補文に対し，「候補識別」器において，N-gram の出現確率に基づく自動識別を適用した場合の効果についての評価を行う．

3. 実験・評価

類似候補文に対し，N-gram(本実験では N=4)を素性として学習した識別器により識別結果文を得，SMT を訓練した．識別器による影響を BLEU 値による客観評価，及び主観評価で確認した．



	1) 原文	2) 類似候補文	3-1) 識別結果文 (正規化無)	3-2) 識別結果文 (正規化有)
ⅰ)1.0Mエントリ	19.68	21.59	21.78	22.17
ⅱ)2.5Mエントリ	19.62	21.38	21.63	21.73

Fig. 4 BLEU score of the evaluation corpus sets

使用したコーパスの詳細を Table 1 に示す．具体的な少量コーパス例として，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(原文)を用いた．この原文に対し換言データベースを適用し類似候補文を生成した．さらに，生成された類似候補文に対し，識別器を適用し識別結果文を得た．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 45 万文対)を加えている．評価文は道案内タスクから約 300 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．

Fig. 4 に(1)原文，(2)類似候補文，(3-1)識別結果文(正規化無)，(3-2)識別結果文(正規化有)を用いた翻訳モデルの性能を示す (BLEU 値)．ベースコーパスに対する，(1)原文・(3-2)識別結果文の文数による影響を避けるため，各々の文数が(2) 識別結果文の文数と略同一となるまで複製し正規化を図っている．各翻訳モデルは各条件において 10 回ずつ生成している．Fig.4 は 10 回の平均値を示している．

3.1 客観評価

得られた BLEU 値について考察する．(1)原文だけの場合に比べて，(2)類似候補文，(3)識別結果文を用いた翻訳モデルを比較すると，ⅰ)1.0M エントリにおいて，平均値で(1)から(2)で+1.91 ポイント，(2)から(3-1)で+0.19 ポイント，(2)から(3-2)で+0.58 ポイントの BLEU 値の改善が見られた．また，ⅱ)2.5M エントリにおいても，平均値で(1)から(2)で+1.76 ポイント，(2)から(3-1)で+0.25 ポイント，(2)から(3-2)で+0.35 ポイントの BLEU 値改善が見られ，良化が確認された．

(1)原文から(2)類似候補文においては，言い換えにより多くの表現を得たことがスコアに寄与していると示唆される．(2)類似候補文から(3-1)識別結果文(正規化無)，(3-2)識別結果文(正規化有)においては，識別器を適用することで“質の良い文”の割合が増え，スコアが上昇したことが考えられる．特に，(2)類似候補文から(3-1)識別結果文でのスコアの上昇においては，学習に用いられる対訳コーパスの総数自体が減少しているにも関わらず，スコアとして改善が見られている．識別器の適用により，ほぼ同等の翻性能で翻訳モデルのサイズを小さくすることができることが示唆された．

Table 2: Translation examples

入力文1	**大阪駅へはどうやっていけばいいんでしょう**
(1)原文	I wonder how can I get to osaka.
(3-2) 識別結果文	**How should I get to osaka ?**
他の翻訳機	It would be nice if we How is to Osaka.
入力文2	**お土産はどこで買ったらいいんですかね**
(1)原文	I'm afraid where can I buy souvenirs..
(3-2) 識別結果文	**Where can I buy that souvenir?**
他の翻訳機	I Where do you say you buy a souvenir.
入力文3	**なんか面白いイベントある？**
(1)原文	There is an interesting events.
(3-2) 識別結果文	**Is there an interesting events?**
他の翻訳機	Softening some interesting events?

Table 3: Examples of verified corpora

識別成功例	(原文) まっすぐに進むと大阪駅です。
	まっすぐに進むと大阪駅**である。**
	まっすぐに**突き進む**と大阪駅です。
	(原文)正面に喫煙所がある。
	正面に**煙草を吸う所**がある。
	前に喫煙所がある。
識別失敗例	(原文) お店に入ります。
	お店へ**曲がれ**。
	(原文) まっすぐに歩いてね。
	ピンとに歩いてね。

以上により，識別素性に N-gram を用いた「候補識別」器の適用は有効であることが示唆された．

3.2 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(1)原文として示している．また，候補識別器を用いて生成された識別結果文での翻訳モデルによる出力結果を(3-2)識別結果文として示している．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．

(1)原文と(3-2)識別結果文の訳文を比較すると，「候補識別」器を用いた識別結果文による(3-2)識別結果文では，比較的良好な翻訳文の出力が確認できる．

また，Table3 に識別結果文の事例を示す．Table3 には，

- 識別器によって良いと識別され，実際に正しかった文(識別成功例)
- 識別器によって正しいと識別されたが，実際には正しくなかった文(識別失敗例)

の例を挙げた．例えば，「喫煙所」⇔「煙草を吸う所」といった単純な言葉の言い換えは比較的精度が高く，一般的に出現頻度の高いフレーズは正しく識別されている．一方で「に入ります」⇔「へ曲がれ」のように，言い換えを行うと元々の文意から離れてしまうが，文法的な誤りのない文が識別結果文として採用されている場合も見受けられる．特に，対訳コーパスにおいては，原言語側の文意を自由に変えられないという制約があるため，例えば，文ベクトルを素性として使うなど，対となる目的言語の文意から離れないような工夫が必要である．

4. おわりに

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．

本稿では「類似候補文生成」器により生成された類似候補文に対し，「候補識別」器において，N-gram の出現確率に基づく自動識別を適用した場合の効果について報告した．特に BLEU 値(平均値)で，約 2.49 ポイントの向上効果を得た．

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] 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)

[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] “statistical machine translation system” http://www.statmt.org/moses/

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

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

[8] 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)

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

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

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