パラレルコーパスからの機械翻訳向け同義表現抽出

下畑光夫[†,††] 渡辺太郎[†]
隅田英一郎[†] 松本裕治[††]

自然言語では同義表現，すなわち主要な意味を共有する表層的に異なる表現，が存在する．この同義表現の多様性は，用例翻訳や統計翻訳といったコーパスに基づく機械翻訳の性能低下を招く．本論文では，パラレルコーパスから同義表現を自動獲得し，それらを利用することによりコーパスに基づく翻訳の性能を向上させる方法について述べる．パラレルコーパスにおいて訳文が等しい文を同義文と考え，同義文間に存在する差異を DP マッチングにより抽出し，さらに頻度に基づいたフィルタリングを行い同義表現を獲得する．獲得する同義表現は，単言語の観点の基で同義である表現だけでなく目的言語へ翻訳した段階で同義である表現も含んでおり，翻訳において有用となっている．さらに，機能語に関する同義表現が獲得できるという特長も備えている．用例翻訳において獲得した同義表現の統一を施したところ，日英，英日の双方向で翻訳可能な文を約 8%拡大し，訳質の低下もわずかにとどまった．また，統計翻訳において同義表現の統一を施したコーパスで学習させたところ，日英の翻訳において訳質が 2.5%向上した．

Extracting Synonymous Expressions from a Parallel Corpus for Machine Translation

Mitsuo Shimohata,[†,††] Taro Watanabe,[†] Eiichiro Sumita[†]
and Yuji Matsumoto[††]

There exists various synonymous expressions, that is, superficially distinct expressions sharing a main meaning, in natural language. The variety of synonymous expressions causes performance degradation in corpus-based machine translation such as example-based machine translation (EBMT) and statistical based machine translation (SMT). In this paper, we propose a method for extracting synonymous expressions from a parallel corpus and utilizing them to improve the performance of corpus-based machine translation. Differences between synonymous sentences which share the same translation are extracted by DP-matching and are then filtered based on frequency to obtain synonymous expressions. Synonymous expressions extracted by our method include those which are synonymous under not only the monolingual viewpoint, but also the translingual viewpoint, and this feature is favorable for machine translation. Moreover, they include synonymous expressions concerning functional words. Synonymous expressions from our method expands translatable inputs of EBMT about 8% without significant degradation of translation quality in J-to-E and E-to-J translation. The performance of SMT was improved 2.5% on J-to-E translation by unifying synonymous expressions in a training corpus.

1. はじめに

自然言語では，表層的に異なるが主要な意味を共有する表現（同義表現）が存在する．この同義表現の多様性は自然言語の表現力の豊かさを示すものであるが，コーパスに基づく機械翻訳においては性能を低下させるという問題がある．

たとえば，用例翻訳はパラレルコーパスから入力文と類似した文を検索しその訳文を修正することによって入力文の翻訳文を生成する翻訳方式である[1)]．入力文とコーパス中の各文について類似か否かを判定する処理において，両文が同じ意味であっても表層的な差異が大きくなると類似文と判定されない．このため類似文を検索する性能が低下してしまい翻訳可能な入力文の減少を招く．

統計翻訳は情報理論で用いられる noisy channel

† ATR 音声言語コミュニケーション研究所
ATR Spoken Language Translation Research Laboratories
†† 奈良先端科学技術大学院大学
Graduate School of Information Science, Nara Institute of Science and Technology

model の概念を翻訳に応用した翻訳方式である[2]．与えられた原言語の文に対して生起確率が最大になるような翻訳文を算出することで翻訳を行う．目的言語文算出のための確率パラメータはパラレルコーパスから学習される．同じ意味を表す事物が同義表現により多様になることで学習コーパス中の言語現象が複雑化し，データスパースネスの問題が大きくなり，翻訳の訳質低下を招く．

本論文では，パラレルコーパスから同義表現を自動獲得し，獲得した同義表現をコーパスに基づく機械翻訳に利用することで性能を向上させる方法について述べる．訳文が同一である文を同義文とし，同義文の間にある差異を単語を単位とした DP マッチングにより抽出し，さらに頻度に基づいた選別を行うことで同義表現を獲得する．獲得には形態素解析されたコーパスを利用し，構文情報や対訳情報といった他の言語的情報を必要としない．本手法で獲得する同義表現には，同義表現の一般的定義である単言語の観点における同義表現に加え，翻訳した段階における同義表現（以後，対訳的観点における同義表現と呼ぶ）も獲得できる．この対訳的観点における同義表現は翻訳という用途において有用となる．さらに，既存のシソーラスや自動獲得手法ではあまり対象となっていない機能語に関する同義表現も獲得できる．そして，本手法で獲得した同義表現を利用することにより，用例翻訳や統計翻訳といったコーパスに基づく機械翻訳の性能を向上させることができる．

以下，2 章で本論文で獲得対象となる同義表現の特徴を述べ，3 章でパラレルコーパスから同義表現を獲得する方法について述べる．4 章では自動獲得した同義表現ならびにそれらを用例翻訳と統計翻訳に利用した実験について述べる．5 章で関連研究について概観する．

2. 獲得する同義表現の特徴

2.1 対訳的観点における同義表現

同義表現は，一般的には単言語の観点で同じ意味を持つ表現を表す．翻訳を目的とした場合にはこの定義による同義表現に加え，目的言語へ翻訳した段階で同義と定義される同義表現も有用と考えられる．

表 1 に対訳的観点における同義表現の例を示す．(1) の 2 つの英文では “wallet” と “purse” という差異がある．これらはそれぞれ「札入れ」と「小銭入れ」という指示物の違いがあり，英語の文化では同一視し難い．しかし，日本語に翻訳する場合には両単語とも「財布」と訳されることが多く，その場合には両者を同一視してもかまわないことになる．(2) の 2 つの和文では “姉” と “妹” という差異があり，日本語の文化では両者を同一視できるとは考えにくい．しかし，英語に翻訳する場合は同じく “sister” と訳されることが多い．これは英語の文化では姉妹の年齢関係は特に訳出することが少ないためである．つまり，表 1 に示している差異は，単言語の観点では同一視できないが日本語/英語に翻訳するという目的のもとでは同一視してもかまわない表現といえる．本手法では対訳文が一致することを文の同義性の拠所としているため，同じ訳になるという対訳的観点における同義表現も獲得できる．

表 1 対訳的観点における同義文
Table 1 Synonymous sentences in translingual viewpoint.

(1)	対訳的観点における同義文	I had my **wallet** stolen? I had my **purse** stolen?
	目的言語文	財布を盗られました
(2)	対訳的観点における同義文	こちらは私の**姉**です こちらは私の**妹**です.
	目的言語文	This is my sister.

2.2 文脈を含んだ表現

ある 2 つの表現が同じ意味を持つかどうかはそれらの表現の周りの文脈に依存することが多い．たとえば，“photos” と “pictures” という 2 単語の同義性を考えてみよう．“photos” は「写真」という意味しかないが，“pictures” には「写真」と「絵画」という 2 つの意味がある．“pictures” に意味的あいまい性があるために，この 2 単語単独では同義か否かは一概には判定できない．そこで，これらの単語に隣接する語を文脈情報として与えると，意味的あいまい性が解消されるために同義性を判定することができる．たとえば，“take pictures” であれば「写真」の意味であるし，“draw pictures” であれば「絵画」の意味となり，“photos” との同義性を判定できる．同様に，“would” と “could” という 2 単語はそれ自体では同義とは考えられないが，前後の文脈を与えた表現 “#☆ (Would | Could) you” においては同義であると判定できる．つまり，同義表現の同義性を考える際には表現単独だけでなく，それに近接する語も文脈情報として合わせて考える必要がある．本論文ではこれ以降，同義性を考える単位となる “表現” は差異である単語列とその近接語を合わせた部分と定義する．

文脈情報として採用する近接語の条件にはその方向と範囲に様々な種類が考えられる．近接語の方向につ

☆ 文頭を表す記号.

いては差異部分の前後双方をとることとした．前後双方をとることで日本語と英語における様々な同義表現に対して共通に適用することができる．ただし両方向をともにとることで条件として不要な近接語を取り込む可能性は高くなる．また，近接語の範囲は両方向 1 語ずつとした．2 語以上を近接語の範囲とすると強い制約となるために同義表現の獲得ならびに利用が大きく減少するためである．4 章で報告するように，1 語だけを文脈情報としても十分な精度を達成している．

2.3 様々な品詞に関わる同義表現

本手法では，内容語はもとより機能語に関する同義表現も獲得することができる．既存のシソーラス（語の意味を階層的に記述した情報）や類義語の自動獲得の研究では機能語は対象外となっている（5 章参照）．しかし，機能語に関わる様々な同義表現は数多く存在しており無視できない．たとえば，日本語においては多様な文末表現が存在しており，依頼表現に限定しても，“しろ”，“してください”，“していただけませんか” などと様々な丁寧さや依頼度を表す文末表現が考えられる．これらは主に助動詞や終助詞などの機能語による差異である．また，助詞についても “東京 (に | へ) 行く”，“日本語 (が | を | の) 分かる人” など様々な同義表現がある．

2.4 語彙的同義表現

本論文では語彙的な差異を持つ同義表現を対象としており，それを超える差異は対象としていない．具体的には，連続 2 単語以内の差異を語彙的差異と定義した．2 単語以内の差異を対象とすることで 2 内容語で構成される複合語や 1 内容語と 1 機能語で構成される文節や 2 機能語で構成される文末表現などに関する同義表現を獲得することができる．なお，語彙的差異を 3 単語とすると非語彙的差異が多く抽出されることが予備実験から判明している．

さらに，語彙的同義表現は，局所的であり他の同義表現と関連しないものと定義した．局所的でない同義表現は英語によく見られる．たとえば “**Are** you **taking** any pills regularly?” と “**Do** you **take** any pills regularly?” の 2 文の間に存在する 2 カ所の差異（ボールド体で記述）は関連しており，それぞれを単独で同義表現として取り出すことはできない．また，語の配置による差異も局所的とはいえず本研究の対象としない．配置が変化する語としては副詞や感動詞などがある．たとえば，“Would you **please** call me a taxi?” と “Would you call me a taxi **please**?” が該当する．

表 2 に日本語における語彙的，非語彙的同義表現の例を示す．図中，文は単語を “_” で区切っている．また，同義文で原文と異なる部分はゴシックで示されている．

表 2 語彙的差異
Table 2 Example of lexical difference.

原文	これ_は_機内_に_持ち込ん_で_も_いい_です_か
語彙的差異を持つ同義文	これ_は_機内_に_持ち込ん_で_も_いい_**でしょう**_か
	これ_は_**飛行機**_**へ**_持ち込ん_で_も_いい_です_か
非語彙的差異を持つ同義文	これ_**を**_**持っ**_**て**_**飛行機**_に_**入っ**_**て**_いい_です_か
	これ_は_機内_**持ち込み**_**でき**_**ます**_か

同義表現には本論文で対象としている語彙的同義表現以外にも差異が及ぶ範囲が句や文など より大きくなるものも考えられる．差異の範囲が様々である同義表現において語彙的同義表現は最小単位である単語を対象としたものであり，より大きな範囲の同義表現の自動獲得を行っていく基盤となる技術であると考えている．

3. 同義表現の獲得

同義表現は，基本獲得処理とその繰返し処理により獲得される．基本獲得処理では，等しい訳文を持つ文を同義文と定義し，同義文間の差異を抽出することで同義表現を獲得する．基本獲得処理で獲得した 2 言語の同義表現を利用し同義表現を統一したうえで同義文集合を形成すると，同義文集合を統合することができる．この統合された同義文集合から再度基本獲得処理を行うことで新たな同義表現を獲得することができる．この処理を繰返し処理と呼ぶ．基本獲得処理については 3.1 節で，繰返し処理については 3.2 節で述べる．

本章では，日英のパラレルコーパスを基に英語の同義表現を獲得する場合を例として説明する．そして，パラレルコーパスを和文 Js_i とそれに対応する英文 Es_i の対からなる集合 $\{(Js_1, Es_1), (Js_2, Es_2), \ldots, (Js_n, Es_n)\}$ として表す．パラレルコーパス中の和文と英文双方において，等しい文が存在しているとする．（例：$Js_1 = Js_4 = Js_{11}$，$Es_1 = Es_5$）

3.1 基本獲得処理

3.1.1 同義文集合の形成

パラレルコーパス中で，同一の対訳文を持つ文を集め，同義文集合を形成する．たとえば，$Js_1 = Js_4 = Js_{11}$ であるとすると，$\{Es_1, Es_4, Es_{11}\}$ が 1 つの同義文集合を形成する．**表 3** に，和文「写真を撮ってもいいですか」により形成された英語の同義文集合の例を示す．

3.1.2 同義表現対の抽出

同義文集合に含まれる複数の同義文からすべての組合せの同義文対を取り出し，そこから同じ意味を持つ

表 3 同義文集合
Table 3 Synonymous sentence group.

共通原文	写真を撮ってもいいですか
同義文（英語）	(1) Can I take pictures? (2) May I take photos? (3) May I take some photos? (4) Can I take a photo? (5) Is it OK to take pitures?

対の同義表現（同義表現対）を抽出する．表 3 の場合，同義文が 5 文あるため 10 通りの同義文対を取り出すことができる．これらの同義文対に以下の処理を施す．

(1) 同義文対のうち，2 文間の差異が小さい同義文対を選別する．
(2) 上記条件を満たす同義文対から 2 文間の差異を抽出し，同義表現対の候補とする．

これらの処理 (1), (2) のための指標として編集操作（edit operation）ならびに編集距離（edit distance）を用いる．編集操作は 2 同義文の一方を他方に変換するために必要となる操作（operation）であり，置換（substitution），削除（deletion），挿入（insertion）の 3 種類がある．編集距離は 2 つの単語列の間における重みつきの編集操作の和を表す．編集距離と編集操作は DP マッチング[3]により算出する．

実験では置換，削除，挿入のいずれも重み 1 とし，編集距離が 2 以下の同義文対を選別すると設定した．編集距離が 3 以上となる同義文対からは，2.4 節で述べた局所的でない非語彙的同義表現が多く抽出されるためである．また，選別された同義文対から得られた編集操作のうち，削除と挿入は除外する．これらの編集操作により獲得される同義表現の多くは “for me” や “today” といった状況に関する語に関する語の削除または挿入であったり，2.4 節で述べた語の配置の違いによるためである．

図 1 に，表 3 中の同義文 (1) と (2) からの同義表現対の抽出を示す．両文の間で DP マッチングを行うと，編集操作として “can” と “may” の置換と “photos” と “pictures” の置換，そして編集距離として置換操作 2 つによる 2 が得られる．編集距離が 2 であることからこの同義文対は獲得対象となる．同義表現対は編集操作とその前後の 1 語を合わせたものであることから，図 1 の下部に示す 2 つの同義表現対が獲得される．

3.1.3 同義表現対のフィルタリング

前項の処理により，各同義文集合から同義表現対が獲得される．これらの同義表現対には不適切なものも多く含まれているため，以下の 2 つの条件により不適切なものを除去する．フィルタリングの条件は以下の 2 つである．これらの条件のしきい値は，予備実験を基に決定した．

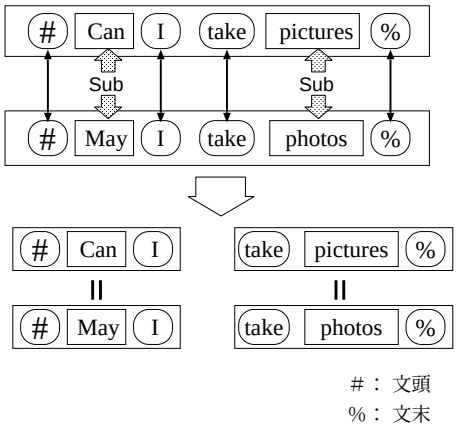

図 1 同義表現対の抽出
Fig. 1 Extraction of synonymous expression pairs.

同義表現対とそれを構成する 2 表現との出現頻度比

“表現 1 = 表現 2” なる同義表現対が候補となっている場合に，表現対そのものの頻度☆と表現対を構成する 2 表現それぞれの頻度の比率を条件に選別する．実験では経験的に以下の式を満たす同義表現対を残すものとした．

$$\frac{freq(\text{“表現 1 = 表現 2”})}{min(freq(\text{“表現 1”}), freq(\text{“表現 2”}))} > 0.05$$

これにより，限られた状況でしか同義性が成立しない同義表現対を除去することができる．

出現した同義文集合の出現頻度

少数の同義文集合にしか出現しなかった同義表現対を除去する．これは学習データ中の不適切な部分に由来する同義表現対を除去するためである．学習データには，単言語部分においては表記誤りにより，2 言語部分においては意訳や文脈に深く依存した訳により不適切な部分が存在する．日本語，英語の双方について出現頻度が 2 回以下の同義表現対を除去するとした．

3.1.4 同義表現集合の形成

前項までの処理により，同義表現対が獲得される．獲得された同義表現 “対” を推移性を利用して同義表現 “集合” にまとめる．つまり，{表現 1 = 表現 2}, {表現 2 = 表現 3} なる 2 つの同義表現対が得られた場合，同義表現集合 {表現 1, 表現 2, 表現 3} が形成される．このようにして得られる同義表現集合とは，

☆ 頻度は出現した同義表現集合の数でカウントする．

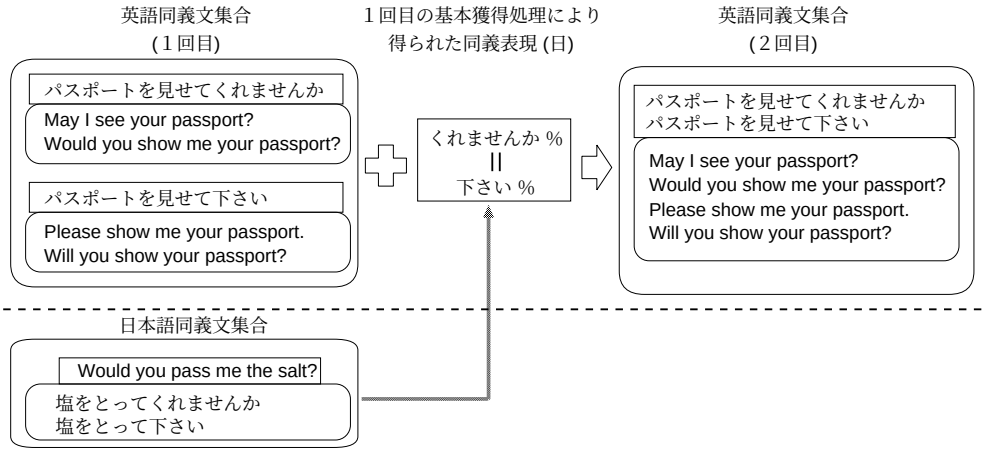

図 2 繰返し処理
Fig. 2 Iteration process.

主要な意味を共有する複数の表現が形成する表現の集合を意味する．

次に，同義表現集合を構成する複数の表現の中から最大の頻度を持つ表現を選出し標準表現とする．最も頻度が高い表現は学習コーパスにおいて最も一般的に使用されるもしくは使用状況が最も広い表現と考えられるからである．獲得した同義表現集合は，非標準表現を標準表現に置換するという形で利用される．

3.2 繰返し処理

3.1 節では日本語の訳文を基にして英語の同義表現を獲得する基本獲得処理について述べた．この日本語と英語の役割を入れ替えて基本獲得処理を行うことにより，日本語の同義表現を獲得することができる．これらの獲得した 2 言語の同義表現を利用して再度基本獲得処理を行うと，新たな同義表現が獲得できる．基本獲得処理の最初に同義文集合を形成するが，その段階で同義表現の統一を行うことで前回の基本獲得処理と異なる同義表現集合が形成されるためである．

図 2 に，同義文集合が繰返し処理により変化する例を示す．1 回目の同義文集合では，訳文が完全に一致するという条件により 2 つの英語の同義文集合が形成されている．1 回目の基本獲得処理により日本語の同義文集合から「くれませんか」と「下さい」という 2 表現が同義表現として抽出される．この同義表現を利用すると，図中の 2 つの英語の同義文集合は同一視できることが分かる．そこで 2 回目の基本獲得処理は，1 回目では別個であった 2 つの同義文集合を 1 つの同義文集合に統合して行う．

このように 1 回目と異なる同義文集合から，新たな同義表現対を獲得することができる．1 回目で獲得した同義表現対と新たに獲得した同義表現対をあわせて同義表現集合を生成し，2 回目の繰返し処理による同義表現集合が獲得される．

同様にして，処理を繰り返していくことで同義文集合を漸増的に獲得する．この繰返し処理は，新たな同義表現が獲得できなくなるまで行う．

4. 実　　験

実験に用いた旅行会話コーパスについて 4.1 節で，そのコーパスから獲得した同義表現について 4.2 節で述べる．そして，用例翻訳と統計翻訳の両翻訳手法について獲得した同義表現を導入した方式と導入しない方式の間で性能を比較した．用例翻訳への適用実験については 4.3 節で，統計翻訳への適用実験については 4.4 節で述べる．

4.1 コーパス

海外旅行でよく用いられる旅行会話を収録した日本語と英語のパラレルコーパスを実験に用いた[4]．このコーパスは日本語と英語が文単位で対応づけられており，形態素解析もされている．このコーパスを学習データと入力データに分割した．学習データは同義表現獲得のためのデータであるとともに，コーパスに基づく翻訳のシステム構築に用いる．**表 4** に学習データの文数（のべ，異なり）と平均文長を示す．このデータには日英双方において同一文が多く存在しており，表 4 によれば日英とも約 30%の文に同一文が存在している．

本手法を適用するうえで，このコーパスは文長が比較的短くかつ類似した文が多く存在するという好条件を備えている．新聞記事などの文長が比較的長い文を対象としたコーパスに対して文を単位として本手法を適用した場合には，ほとんど同義表現が得られないと

表 5 獲得した同義表現集合
Table 5 Extracted synonymous expression groups.

同義表現の言語	日本語			英語		
	1 回目	10 回目	拡大率 (%)	1 回目	10 回目	拡大率 (%)
獲得同義表現集合数	1,110	1,251	12.7%	794	980	23.4%
平均表現数/同義表現集合	2.39	2.48	3.8%	2.28	2.36	3.5%

表 4 同義表現獲得に用いたコーパス
Table 4 Corpus for synonymous expression extraction.

	日本語	英語
のべ文数	127,110	127,110
異なり文数	89,615	85,839
平均文長（単語数）	7.8	6.7

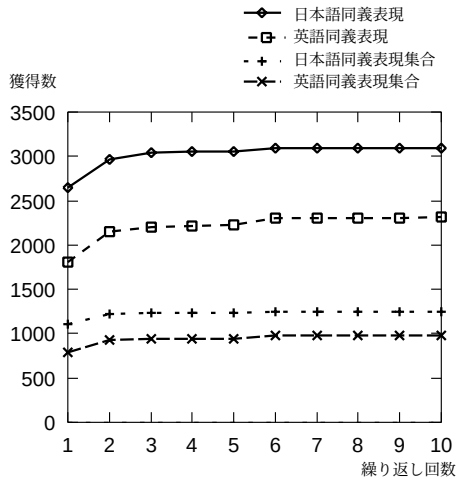

図 3 繰返しによる同義表現抽出
Fig. 3 Extracted synonymous expressions by iteration.

日本語同義表現集合		
いい	です	か
いい	の です	か
いい	でしょう	か
いい	ん です	か
いい	の でしょう	か
#	料金	は
#	値段	は
#	運賃	は
日本語	の	話せる
日本語	を	話せる
日本語	が	話せる
は	どこ	です
は	どちら	です
は	どの 辺	です

英語同義表現集合		
#	Could	you
#	Would	you
#	Can	you
#	Will	you
#	Can't	you
#	Can	I
#	Could	I
#	May	I
a	terrible	headache
a	severe	headache
a	gift	for
a	souvenir	for
am	terribly	sorry
am	really	sorry

図 4 獲得した同義表現集合の例
Fig. 4 Example of extracted synonymous expression groups.

表 6 獲得した同義表現の品詞の内訳
Table 6 POS distribution of extracted synonymous expressions.

日本語		英語	
品詞	割合	品詞	割合
名詞	42.4%	名詞	47.3%
動詞	28.0%	動詞	21.6%
形容詞	3.3%	形容詞	9.3%
副詞	2.2%	副詞	8.3%
助詞	19.0%	前置詞	5.1%
助動詞	22.4%	助動詞	9.7%

考えられる．そのようなコーパスに対しては節や句といった小単位の 2 言語対に対して本手法を適用することで多くの同義表現が得られると考えられる．なお，パラレルコーパスから小単位の対訳対を自動的に求める研究としては，文献 5)〜8) があげられる．

4.2 獲得した同義表現

同義表現獲得用コーパスに基本獲得処理とその繰返し処理を施し，日本語と英語双方の同義表現を獲得した．**図 3** に繰返し処理の回数による獲得した同義表現集合数ならびにそれらに含まれる同義表現の総数を示す．繰返し処理では 10 回目までは新たな同義表現を獲得することができ，11 回目以上からは変化がなかった．

表 5 に 1 回目ならびに 10 回目で獲得した同義表現集合の集合数，1 集合あたりの平均表現数，そして繰返し処理による獲得数拡大率を示す．拡大率より，繰返し処理はすでに得られた同義表現集合内に新たな表現を加えるというよりは新たな同義表現集合を獲得する方の効果が高いことが分かる．また，日本語の方が獲得した同義表現集合，平均表現数とも若干多くなっている．獲得した同義表現の一部を**図 4** に示す．図中の同義表現集合で最上段の表現は標準表現を表している．

本手法では 2.3 節で述べたように様々な種類の同義表現を獲得することができた．**表 6** に獲得した同義表現の主な品詞ごとの割合を示す☆．既存のシソーラスや同義語の自動獲得では機能語の情報は少ないが，本手法では日本語と英語ともに多く獲得できている．

この品詞の割合は言語による特徴も示している．た

☆ 同義表現集合には複数の品詞が関わるものがあるため，各品詞の合計は 100%を超える．

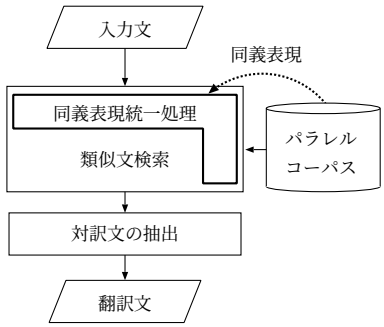

図 5 実験に用いた用例翻訳
Fig. 5 Experimental EBMT system.

とえば，日本語は助詞，助動詞に関する同義表現が高い比率となっている．日本語では助詞は様々な置き換えや省略が可能であり，獲得した同義表現集合にもそれを反映したものが多く獲得されている．助動詞は文末表現で中心的役割を果たしており，様々なレベルの丁寧表現が同義表現として多く獲得された．

また，獲得した同義表現を既存のシソーラスである分類語彙表と比較した．獲得した同義表現から隣接する文脈の情報を除いた差異の部分を対象とした．差異の部分の表現は異なりで 1,584 表現獲得しているが，このうち，分類語彙表に記載されていない表現が 976（61.6%）あった．内容語 1 語で構成される表現に限定しても，獲得した 873 表現のうち，327 表現（37.5%）は分類語彙表に存在していなかった．これより，本手法により分類語彙表にない多くの表現が本手法により獲得できていることが分かる．

4.3 用例翻訳への効果

4.3.1 実 験 環 境

用例翻訳には，原言語側の文に対して同義表現を標準表現に統一するという形で適用した．これにより，入力文とパラレルコーパス中の用例文との間で同義表現の差異を無視し，用例翻訳で翻訳可能な文を拡大することができる．実験に用いた用例翻訳の構成を**図 5** に示す．用例翻訳では入力文と各用例文が類似か否かを判定するが，その際に獲得した同義表現を統一してから類似性を判定する方式（統一あり）と統一しないで判定する方式（統一なし）の 2 方式の性能を比較した．統一なし方式は，入力文と完全一致する用例文を検索してその対訳文を修整なしで出力することで翻訳を行った．統一あり方式は，入力文ならびに用例文に存在する同義表現を標準表現に置換したうえで完全一致する用例文の対訳文を出力することで翻訳を行った[9]．たとえば「写真を撮ってもいいでしょうか」が入力文として，「写真を撮ってもいいのですか」がパラレルコーパスに存在していた場合，両者はそのままでは完全一致しないが標準表現に置換することにより両者とも「写真を撮ってもいいですか」となり，翻訳が可能となる．

統一処理による一致文が複数存在した場合，統一処理前の用例文の段階で入力文の間と最小の編集距離を持つ文を採用する．さらに編集距離最小の文が複数存在する場合は，その中からランダムに 1 文を選択し採用する．

表 7 同義表現統一による翻訳可能文拡大
Table 7 Expansion of translatable inputs by unifying synonymous expressions.

翻訳方向	入力文数	翻訳された入力文数		カバレッジ
		統一なし	統一あり	拡大率 (%)
日英	8,092	3,201	+280	8.7%
英日	7,564	2,890	+244	8.4%

表 8 訳質評価
Table 8 Evaluation of translation quality.

		統一なし	統一あり
日英	評価文数	500	280
	正訳文	486	267
	精度 (%)	97.2%	95.4%
英日	評価文数	500	244
	正訳文	492	239
	精度 (%)	98.4%	98.0%

4.3.2 結　　果

用例翻訳に対する同義表現統一の効果は，翻訳可能文の拡大と新たに獲得した翻訳文の訳質という 2 点で評価した．翻訳可能な文の拡大率は，統一なし方式で翻訳できた入力文に対する統一あり方式により新たに翻訳できた入力文の割合で表される．結果を**表 7** に示す．日英方向では，統一なし方式では 8,092 文の入力文のうち 3,201 文翻訳することができた．統一あり方式により翻訳可能な入力文が 280 文増加したため，翻訳可能文拡大率は 8.7%となる．同様に，英日方向では拡大率は 8.4%となった．

また，目的言語を母語とする評価者により翻訳文の訳質を評価した．表 7 で示した結果のうち，統一なしについてはランダムにサンプリングした 500 文を，統一ありについては追加翻訳可能文全部について訳質を評価した．統一なし方式は入力文と完全一致する文の対訳文を出力しているが，不適切もしくは文脈に大きく依存している対訳文が存在するために，その訳質は 100%ではない．

評価結果を**表 8** に示す．日英，英日とも，統一あり方式は統一なし方式と比較すると日英で 1.8%，英日で 0.4%の訳質低下が見られるものの，統一なしで

表 9 統計翻訳における同義表現統一による訳質の差異
Table 9 Difference of translation quality on SMT by unifying synonymous expressions.

	ランクの割合 (%)			
	A	B	C	D
統一なし	29.2%	23.8%	17.1%	30.0%
統一あり	27.9%	28.8%	15.8%	27.5%

97%を超える精度と比較するとその低下幅は小さいといえる．

4.4 統計翻訳への効果

統計翻訳は，与えられた入力文に対する生起確率が最大となる翻訳文を算出することで翻訳を行う．生起確率は，ベイズの定理に基づき翻訳モデル（目的言語文から入力文が導かれる確率モデル）と言語モデル（目的言語文の尤度モデル）の積に分解される．学習コーパスとして与えられたパラレルコーパスから翻訳モデルと言語モデルのパラメータを学習する．実験では，翻訳モデルとして IBM モデル 4 [2] を用いた．

統計翻訳へは，学習コーパスの目的言語の文に対して同義表現を標準表現に統一するという形で適用した．これにより，目的言語側の表現の多様性が減少するために，翻訳モデルのデータスパースネスを軽減することができる．学習コーパスをそのまま用いて学習させた方式（統一なし）と，学習コーパスの目的言語側の文に同義表現統一処理を施したうえで学習させた方式（統一あり）の 2 方式について性能を評価した [10]．

評価文として日本語文を 240 文与え，出力された翻訳文を人手により A（perfect），B（fair），C（acceptable），D（nonsense）の 4 つのランクに評価した．**表 9** に統一なしと統一ありによる訳質の差異を示す．ランク A，B，C の訳を正解訳と考えると，同義表現統一により正解訳出力率が 2.5%向上しており，同義表現統一が統計翻訳の性能向上をもたらしたといえる．より細かく見ると，ランク B が増加し，ランク A，C，D が減少している．ランク C，D が減少したことは，同義表現統一により統計モデルが変化し性能が向上したことを示している．ランク A は 1.3%減少しているが，この理由としては同義表現の統一により細かい意味やニュアンスが打ち消されてしまったことが考えられる．

5. 関 連 研 究

本章では本論文が対象としている語彙的同義表現に関連する研究について概観する．人手で作成されたシソーラスについて 5.1 節で述べ，語彙的同義表現を自動獲得する研究について 5.2 節で述べる．

5.1 シソーラス

同義語や類義語の情報を階層的に記述したシソーラスには人手で作成されたものが多く存在している．日本語では日本語語彙大系 [11]，分類語彙表 [12]，角川類語辞典 [13] が，英語は WordNet [14] と Roget's Thesaurus [15] がよく知られている．

これらの既存のシソーラスをコーパスに基づく機械翻訳に適用する場合の問題点は，機能語に関する情報が非常に少ないことである．たとえば，日本語語彙大系では約 40 万語について意味体系の情報を記載しているが，機能語である付属語（助詞，助動詞），補助動詞には意味属性が付与されていない．また，分類語彙表に収録されている約 32,600 語のほとんどは名詞（体の類），動詞（用の類），形容詞（相の類）であり，それ以外の語（その他）については 362 語しか記載されていない．コーパスに基づく機械翻訳では機能語の部分も類似文の判定に関わってくるため，機能語に関わる同義表現も必要となってくる．

さらに，既存のシソーラスに記述されている情報は一般的な語，用法にとどまっているため，特定分野の機械翻訳に利用した場合に十分にカバーされるとは限らない．本手法はコーパスそのものから同義表現を獲得するため，分野固有の同義表現が獲得できる．また，対訳的観点の同義表現と語の削除による同義表現を含んでいることも提案手法の特長である．

5.2 語彙的同義表現の自動獲得

単言語コーパスを用いて類義語を自動獲得する研究が多く行われている．既存のシソーラスと比べると分野固有の知識が獲得できる点は解決されているが，機能語に関する同義表現がほとんど獲得対象となっていない問題は残っている．これらの研究の多くは 2 語の類似関係を判定するためにそれら 2 語の使用状況を利用している．ところが，機能語は内容語に付随して内容語の意味を補足するものであり機能語単独の使用状況から類義関係が判定できないため，機能語の類義語獲得が困難なのである．なお，2 語の類義関係を測る指標としては，それぞれの語が現れる文書の共通性や他の語との修飾もしくは被修飾関係の共通性が用いられることが多い [16]．たとえば Lin [17] は係り受け関係にある 2 語とその関係による 3 つ組を利用し，3 つ組を多く共有する 2 語を類義語として獲得している．獲得は名詞，動詞，形容詞に限定している．山本 [18] も同様の手法と既存のシソーラスによるフィルタリングにより類義語より制約が強い同義語を獲得している．さらに，これらの単言語コーパスのみに基づいて獲得する研究では対訳的観点における同義表現は獲得でき

ない．

Barzilay ら[19]は本研究と同じくパラレルコーパスを用いて同義表現（パラフレーズ）を獲得している．パラレルコーパスを用いると文の同義性の情報が得られるため，単言語コーパスのみを用いる場合と比べて多様な情報が獲得できる．彼らは統語的，句レベル，1 語の語彙的同義表現が獲得できると述べている．同義表現の用途として複数文書要約や文生成をあげているが，その具体的効果については述べられていない．

6. ま と め

ある事物を様々な表現で表すことができるという同義表現の多様性は自然言語の表現力の豊かさを示すものであるが，この多様性は自然言語処理において大きな問題を引き起こす．用例翻訳や統計翻訳といったコーパスに基づく機械翻訳においても翻訳性能の低下を招いてしまう．

本論文ではパラレルコーパスから語彙的同義表現を獲得する方法を提案した．等しい訳文を持つ同義文の間で DP マッチングにより差異を抽出し，近接語と合わせて同義表現を獲得した．この手法では，対訳的観点における同義表現も獲得可能であり，翻訳という用途に適したものとなっている．また，内容語と機能語の双方について同義表現を獲得することができた．

そして，獲得した同義表現を用例翻訳と統計翻訳に利用することで翻訳性能を向上させることができた．用例翻訳では日英，英日の双方向において翻訳可能文を約 8%拡大することができた．しかも，同義表現統一により獲得した翻訳文の訳質の低下はわずかにとどまった．また，統計翻訳でも学習コーパスに同義表現統一を施すことにより正解訳出力率を 2.5%上げることができた．本論文で提案したパラレルコーパスからの同義表現の自動獲得はコーパスに基づく機械翻訳の性能向上に有効であるといえる．

謝辞 本研究は通信・放送機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである．

参 考 文 献

1) Nagao, M.: A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, *Artificial and Human Intelligence*, pp.173–180 (1981).
2) Brown, P.F., Cocke, J., Pietra, S.D., Pietra, V.J.D., Jelinek, F., Lafferty, J.D., Mercer, R.L. and Roossin, P.S.: A Statistical Approach to Machine Translation, *Computational Linguistics*, Vol.16, No.2, pp.79–85 (1990).
3) Cormen, T.H., Leiserson, C.E., Rivest, R.L. and Stein, C.: *Introduction to Algorithms*, MIT Press (2001).
4) Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H. and Yamamoto, S.: Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World, *Proc. 3rd LREC*, pp.147–152 (2002).
5) Matsumoto, Y., Ishimoto, H. and Utsuro, T.: Structural Matching of Parallel Texts, *31st Annual Meeting of the ACL*, pp.23–30 (1993).
6) 北村美穂子，松本裕治：対訳コーパスを利用した対訳表現の自動抽出，情報処理学会論文誌，Vol.38, No.4, pp.727–735 (1997).
7) Watanabe, H., Kurohashi, S. and Aramaki, E.: Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Translation, *Proc. 18th International Conference on Computational Linguistics* (*COLING 2000*), pp.906–912 (2000).
8) 今村賢治：構文解析と融合した階層的句アライメント，自然言語処理，Vol.9, No.5, pp.23–42 (2002).
9) Shimohata, M. and Sumita, E.: Automatic Paraphrasing Based on Parallel Corpus for Normalization, *Proc. 3rd International Conference on Language Resources and Evaluation* (*LREC*), pp.453–457 (2002).
10) Watanabe, T., Shimohata, M. and Sumita, E.: Statistical Machine Translation on Paraphrased Corpora, *Proc. 3rd International Conference on Language Resources and Evaluation* (*LREC*), pp.1954–1957 (2002).
11) 池原 悟，宮崎正弘，白井 諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林 良彦（編）：日本語語彙大系，岩波書店 (1997).
12) 国立国語研究所（編）：分類語彙表，秀英出版 (1964).
13) 大野 晋，浜西正人（編）：角川類語新辞典，角川書店 (1991).
14) Fellbaum, C.: *WordNet: An Electronic Lexical Database*, MIT Press (1998).
15) Roget, P.: *Roget's International Thesaurus*, Thomas Y. Crowell (1946).
16) Manning, C.D. and Schütze, H. (Eds.): *Foundations of Statistical Natural Language Processing*, MIT Press, chapter Lexical Acquisition, pp.265–314 (1999).
17) Lin, D.: Automatic Retrieval and Clustering of Similar Words, *COLING-ACL*, pp.768–774 (1998).
18) 山本和英：テキストからの語彙的換言知識の獲得，言語処理学会第 8 回年次大会，pp.639–642

(2002).
19) Barzilay, R. and McKeown, K.R.: Extracting Paraphrases from a Parallel Corpus, *Proc. 39th Association for Computational Linguistics* (*ACL*), pp.50–57 (2001).

(平成 15 年 6 月 2 日受付)
(平成 15 年 9 月 5 日採録)

下畑　光夫（学生会員）

1989 年大阪大学基礎工学部制御工学科卒業．1991 年同大学院基礎工学研究科物理系修士課程修了．同年沖電気工業（株）入社．2000 年より ATR 音声言語コミュニケーション研究所に出向，2002 年奈良先端科学技術大学院大学情報科学研究科博士後期課程入学，現在に至る．自然言語処理の研究に従事．言語処理学会会員．

渡辺　太郎

1994 年京都大学工学部情報工学科卒業．1997 年同大学院工学研究科情報工学専攻修士課程修了．同年同大学院工学研究科情報工学専攻後期博士課程入学．2000 年 Language and Information Technologies，School of Computer Science，Carnegie Mellon University，Master of Science 取得．2001 年より ATR 音声言語コミュニケーション研究所研究員，現在に至る．言語処理，機械翻訳，統計的機械翻訳の研究に従事．言語処理学会会員．

隅田英一郎（正会員）

1982 年電気通信大学大学院計算機科学専攻修士課程修了．1999 年京都大学工学博士．現在，ATR 音声言語コミュニケーション研究所主任研究員．工学博士．自然言語処理（特に，機械翻訳，統計的文脈処理，音声言語統合処理，言語データベース，並列自然言語処理，情報検索）の研究に従事．電子情報通信学会，言語処理学会，ACL 各会員．

松本　裕治（正会員）

1977 年京都大学工学部情報工学科卒業．1979 年同大学院工学研究科修士課程情報工学専攻修了．同年電子技術総合研究所入所．1984 年〜85 年英国インペリアルカレッジ客員研究員．1985 年〜87 年（財）新世代コンピュータ技術開発機構に出向．京都大学助教授を経て，1993 年より奈良先端科学技術大学院大学教授，現在に至る．工学博士．自然言語処理の研究に従事．人工知能学会，日本ソフトウェア科学会，言語処理学会，認知科学会，AAAI，ACL，ACM 各会員．