社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2006－NL－176（7）
2006／11／22

対訳文選択のための用例翻訳用シソーラスの構築

福井 健司† 　柏岡 秀紀†‡§

†奈良先端科学技術大学院大学情報科学研究科
‡独立行政法人情報通信研究機構
§ATR 音声言語コミュニケーション研究所
E-mail: †{kenji-fu, kashioka}@is.naist.jp

近年, 機械翻訳はコーパスベース翻訳が注目されており, 我々は用例翻訳に焦点を当て研究を行っている. 用例翻訳は入力文の類似用例を用例データベースから取得し, それらを編集することで翻訳を行う. この類似用例取得にシソーラスがよく利用される. しかし, 既存のシソーラスは単一言語のみを考慮して構築されたものがほとんどであり, そのようなシソーラスは多言語間での概念相違を考慮しておらず, 又, 概念体系も固定的であり, 結果として翻訳に正しくない用例の取得や翻訳に有効な用例の取りこぼしなどを引き起こし, 翻訳の精度を落としかねない要因になる. 本稿ではそのような問題解決のために, 単一言語コーパスからではなく対訳コーパスを利用し多言語情報からのシソーラス構築を提案する. 本稿では構築手法を示し, 既存のシソーラスと構築したシソーラスのそれぞれを用いて用例翻訳を行なった実験結果について報告する.

Building A Thesaurus Based on Example-Based Machine Translation for Selecting A Translation Pair from Example Data Base

Kenji Fukui† 　Hideki Kashioka†‡§

†Graduate School of Information Science, Nara Institute of Science and Technology
‡National Institute of Information and Communications Technology
§ATR Spoken Language Communication Research Laboratories
E-mail: †{kenji-fu, kashioka}@is.naist.jp

In machine translation research, corpus-based translation has been a focus of attention. We focus on Example-Based Machine Translation (EBMT). For a basic idea of EBMT, EBMT retrieves some examples that are similar to an input sentence and adjust them to obtain the translation. For retrieving some examples, a thesaurus is commonly used. But most existing thesauri are usually built with monolingual knowledge and they do not consider the conceptual differences between languages and they usually have a fixed conceptual system. To use such thesaurus would reduce the translation accuracy by retrieving wrong examples and not retrieving available examples. In this paper, to solve such problems, we propose a thesaurus that is built with a parallel corpus to use bilingual knowledge, instead of a monolingual corpus. We show the building method and report experimental results of EBMT by using an existing thesaurus and the proposed thesaurus in a comparison.

1　はじめに

近年, 機械翻訳は大量の対訳文 (対訳コーパス) を用いた翻訳手法 (コーパスベース翻訳) が注目されており, 現在では統計翻訳 [1] と用例翻訳 [2] の２つの手法が活発に研究されている. 本稿は用例翻訳に焦点を当て研究を行なっている. 用例翻訳の基本的な考え方は, 入力文と類似している用例を選択し, それらを編集することで翻訳を行なう方式である. この考え方に基づいて, これまで多くの翻訳手法が報告されている [3][4][5][6].

類似用例の選択基準は文単位もしくは節単位など翻訳手法によって異なるが, どちらの基準にしても用例データベースからいかにして有効な用例を選択するかが重要である. 例えば”**シーディー**をかける”という簡単な入力文を翻訳する場合でも,” **レコード**をかける (play a record)”, ” **お金**をかける (bet money)”,” **目覚まし**をかける (set an alarm)”,” **エンジン**をかける (start an engine)”のように「**X** をかける」という表

現は多数あり，　～　の中から入力文の翻訳に適切な用例を選択をしなければ翻訳は失敗する. そこで用例翻訳では適切な用例選択のためにシソーラスがよく利用される. シソーラスは単語の概念を体系（上位/下位関係, 部分/全体関係, 同義関係, 類義関係等）付けた辞書であり, シソーラスを利用することで単語間の意味的距離 (概念距離) を求めることができる. シソーラスを上述の例に適用すると　が入力文の翻訳に適切 (”シーディー”と”レコード”の概念は等しい. 表 1 参照) な用例であり, この用例を類似用例として選択することで正しい翻訳に導くことができる.

表 1:単語 A,B 間の概念

単語 A	単語 B	概念
シーディー	レコード	○
シーディー	お金	×
シーディー	目覚まし	×
シーディー	エンジン	×

※同一概念→○, 異なる概念→×

しかし, 既存のシソーラスを利用することで翻訳の精度を落としかねない事例が存在する. 以下に例を示す.

問題　共通概念だが翻訳に失敗する事例
入力文：電車 に**乗る**(正解：”**take** a train”)
対訳文 1：”自転車 に**乗る**”－”**mount** a bicycle”
対訳文 2：”バス に**乗る**”－”**take** a bus”

下線部の単語は「乗り物」という共通概念を持ち, 各対訳文は類似用例とみなされるが, 対訳文 1 は入力文の翻訳とは微妙に表現がずれ (take と mount), もし対訳文 1 を類似用例として選択すると翻訳に失敗する.

問題　異なる概念だが翻訳に有効な事例
入力文：食欲 がない (正解：”I have no appetite”)
対訳文 3：”小銭 がない”－”**I have no** small change”

下線部の単語は一般的には異なる概念として認識され類似用例としてみなされないが, 対訳文 3 の英語表現 (I have no X) は入力文の翻訳に有効である.

以上の問題点を整理すると

問題　：　翻訳側での表現のずれを招く恐れ
問題　：　有効な用例の取りこぼし

となる. これは既存のシソーラスおよび従来のシソーラス構築の研究は, そのほとんどが単一言語を対象としており, その結果, 多言語間での概念相違 (問題　の原因) を考慮していないことや概念体系が固定的 (問題　の原因) であることが, 上述の問題を引き起こす要因であると考えられる.

本研究では, 上述の問題点を改善をすべく, 既存の概念体系とは異なるシソーラス, すわなち機械翻訳に適したシソーラスの構築を提案する. 例えば上述の例題の場合, 表 2 に示す「本研究が目指す概念」が構築できれば, 上述の問題解決が期待できると考えている.

表 2:単語 A,B 間の概念 (例題より)

単語 A	単語 B	既存の概念	本研究が目指す概念
電車	自転車	○	×
電車	バス	○	○
食欲	小銭	×	○

※同一概念→○, 異なる概念→×

本稿では,2 節で本研究のベースとなるシソーラス構築を紹介すると共に本研究における構築の方向性について述べる.3 節ではシソーラス構築について述べる.4 節では構築実験と評価実験について述べ, それぞれの結果について考察する. 最後に 5 節でまとめと今後の課題について述べる.

2　関連研究

本研究では [7] の手法をベースにしている.[7] は言語解析 (形態素解析や構文解析等) が行なわれていない単一言語コーパスから同義関係の語 (これを関連語と呼ぶ) を自動的に獲得することを行なっている. はじめに, 関連語候補となる語の対をテキストから自動的に抽出し, 抽出された候補の関連性を図るスコア関数を設定し, このスコア関数の値に基づいて関連語を獲得している. スコア関数は関連語対の前後文字列を判定基準としている. 例えば”プリント”－”印刷”という候補を考えた場合,「年賀状を (印刷｜プリント) しなければならない」という文はどちらの単語を選択しても文意は同じである. このように同義関係の語の対を判定する場合,「前後の文字列が同じように使用されている」という事象は非常に有効な手掛かりである.[7] はこの事象に $tf \cdot idf$ を組み込むことでスコア関数を設定し, 関連語の判定を行なっている.

本研究で獲得すべきものは表 1,2 に示すような単語対とその概念である. ある単語対が類似用例選択に有効であるかを判定する手掛かりとして「両単語 (入力文側と用例側) の周辺文脈が同じように使用されている」という事象は [7] と同じく有効な手掛かりとなる. しかし, この事象を単一言語のみに適用するだけでは既存のシソーラスとなんら変わりがない. そこで本研究では, 単一言語コーパスではなく対訳コーパスを利用することで, 上述の事象を原言語 (翻訳元) だけでなく目的言語 (翻訳先) の文脈情報も利用することでシソーラスの構築を試みる. 対訳文を利用することで, 問題　のように目的言語側での表現のずれを考慮することができ, 又, 周辺文脈が同じという観点では問題　も同様に考慮でき, 類似用例選択に適したシソーラスが期待できる. 又, 対訳コーパスから構築しているという観点から, 単一言語のみでなく多言語シソーラスの獲得にも発展することが可能である.

本稿で構築するシソーラスのドメインは海外旅行会話に設定し, その対訳コーパスとして BTEC(Basic Trabel

Expression Corpus)[8] の日英対訳文を利用する. 又, これまで示してきた例のように類似用例を選択する際に考慮される単語は名詞に関するものが多いと考えられ, 本稿のシソーラスの対象は名詞に関する単語 (普通名詞, 固有名詞, サ変名詞など) を対象とする.

3 用例翻訳用シソーラスの構築

本稿で提案する手法は, 対訳コーパスから単語対訳対 (これを関連語と呼ぶ) を生成する関連語候補生成部と獲得された関連語候補の関連性を求めて判定を行なう関連語候補判定部の 2 つの処理で構成される. 構築の流れを図 1 に示す. 各処理の詳細は次節以降で説明する.

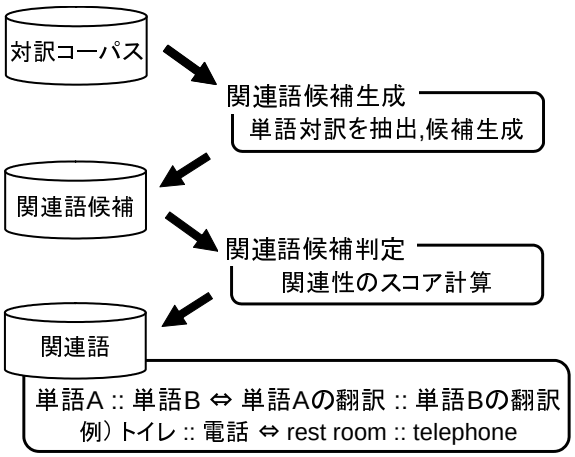

図 1: 構築の流れ

3.1 関連語候補生成部

関連語候補生成部では, 類似用例選択に有効な単語対 (原言語側) とその単語対の翻訳 (目的言語側) を生成し, それらを候補として扱う. 関連語候補生成の指標として単語の共起情報などが一般的に利用されるが, 本稿では BTEC の特徴を利用して候補を生成することにする.BTEC の特徴として, 対訳文が発話単位であるため, 両言語に関して文長が比較的短いことが挙げられる. 表 3 に日英の文長を示す. 文長が短いことに関連して,BTEC では類似文型が得られやすく, その例として,

「X があります」(文長=4), 「X はどこですか」(文長=5), 「X が欲しいのですが」(文長=6), が挙げられる. 上述の例は日本語側だけでなく英語側でも 「I have a X」(文長=4), 「Where is the X」(文長=4), 「I want a X」(文長=4) と同様に類似文型の場合が存在する. 複数の対訳文が, 両言語で同様の類似文型を持つことは翻訳パターンを獲得していることを意味し, その置換部分にあたる単語対 (日本語側の対と英語側の対の両方) は類似用例選択に有効な単語対候補として捉えることができる. 本稿ではこの情報を基に関連語候補を生成する. 以下にその生成手順を示す (図 2 にその詳細を示す).

生成手順

Step1

対訳コーパスから日本語側で 1 単語のみ異なる対訳文を収集する. 同時に, 対訳文内で異なり部分の単語の翻訳を対訳辞書 (本稿で使用する対訳辞書は日本語で約 12 万語登録) で獲得する.

Step2

Step1 で獲得された対訳文を両言語で単一化を行い, 翻訳パターンと単語対訳のリストを作成する.

Step3

Step2 の翻訳パターン内での単語対訳の組み合わせを関連語候補とする.

表 3:BTEC の日英の文長

	日本語	英語
文長	6.50	5.78

※両言語とも文長は形態素単位を考慮

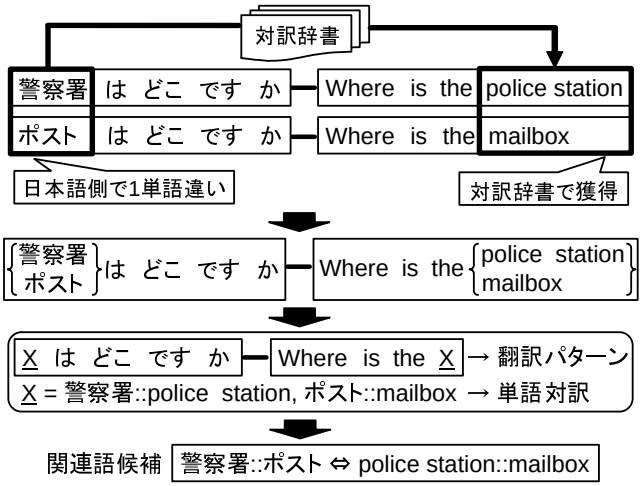

図 2: 関連語候補の生成手順

3.2 関連語候補判定部

関連語候補判定部では,3.1 節で得られた関連語候補の関連性を求め, 関連語かどうかを判定する. 関連性は関連語の各言語での前後の文脈情報を指標とする. ここで関連語の日本語側を A,B, 英語側を a,b(関連語 A::B ⇔ a::b として存在), 日本語側の周辺文脈を X,Y, 英語側の周辺文脈を x,y とし,「XAY::xay」(以後,Z_{Aa}), 「XBY::xby」(以後,Z_{Bb}) が判定に必要な単語列となる.

尚, この単語列は対訳コーパスの各対訳文から抽出されたものである (図 3 参照).

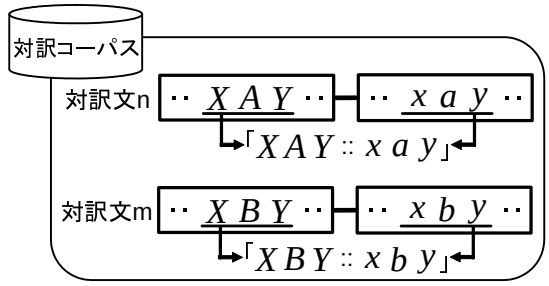

図 3: 対訳文からの単語列の抽出

獲得された単語列を基にして関連性の判定に必要な単語列のスコア関数を式 (1) に示す.

$$score(Z_{Nn}) = tf(Z_{Nn}) \cdot LW(Z_{Nn}) \qquad (1)$$

(但し,$Z_{Nn} = Z_{Aa}$ or Z_{Bb})

式 (1) は単語列の出現頻度と単語列の重みを考慮したものである. ここで各項について説明する.

- $tf(Z_{Nn})$ **について**
  $tf(Z_{Nn})$ は対訳コーパス中における単語列 Z_{Nn} の出現頻度を考慮したものであり, この値が大きいほど関連性の判定に有効に効く指標である. 但し稀な情報を防ぐために $tf(Z_{Nn})$ >1 の制約を入れた. 尚,[7] は tf と共に idf の指標を組み込んでいるが, 本研究で使用する BTEC は発話単位の対訳文集合でありドキュメントを特定しにくいため idf は除外した.

- $LW(Z_{Nn})$ **について**
  $LW(Z_{Nn})$(LW=$Length\ Weigth$) は, 単語列が抽出された対訳文の文長 (これを元文長と呼ぶ) が長いほど関連語の判定に有効に効くという想定の下で考慮した指標である. 具体的には対訳コーパス上での各言語における平均文長と元文長を考慮し, 式 (2) で求める.

$$LW(Z_{Nn}) = \frac{lw(S_{Z_{Nn}}) + lw(T_{Z_{Nn}})}{2} \qquad (2)$$

$$lw(M_{Z_{Nn}}) = exp(1 - \frac{length(M_{mean})}{length(M)}) \qquad (3)$$

S:翻訳元言語, T:翻訳先言語
$M_{Z_{Nn}}$:文字列 Z_{Nn} の元文長 (M=S or T)
$length(M_{mean})$:M の平均文長
$length(M_{Z_{Nn}})$:M の文長

$LW(Z)$ は各言語で元文長が平均文長以上ならば重みを 1, それ以外の場合は元文長が短くなるにつれ 0 に近づく重みとした.

式 (1) の単語列のスコアを基に関連語の関連性をはかる $relate\_score(A :: B \Leftrightarrow a :: b)$ を式 (4) に示す.

$$relate\_score(A :: B \Leftrightarrow a :: b) = \sum_{X,Y,x,y} (score(Z_{Aa}) \cdot score(Z_{Bb})) \qquad (4)$$

式 (4) は関連語の前後文脈は等しいという制約を入れ, この制約下で値が大きいほど関連性が高いことを意味する. 又, この値を閾値操作することで有効な関連語を獲得することができる. 閾値決定の詳細は次節の実験で述べる.

4 実験

4.1 実験で使用する用例翻訳

本稿で使用する用例翻訳は [3] に基づいたものである. 以下に翻訳手順と翻訳評価について説明する (本稿は日本語から英語への翻訳である).

4.1.1 翻訳手順

4.1.1_a 類似用例文取得

[3] は対訳コーパスを用例データベース (各対訳文を用例とする) として扱い, 入力文との類似用例文の取得は両文の間で DP マッチングを計算し, 式 (5) のモデルを用いて行う.

$$dist = \frac{I + D + 2\sum SEMDIST}{L_{input} + L_{example}} \qquad (5)$$

式 (5) の I と D は DP マッチングの挿入操作と削除操作 (両操作の重みは 1 とする) のそれぞれの総和を表している.L_{input} は入力文の文長,$L_{example}$ は用例文の文長 (文長は形態素単位) を表している.$SEMDIST$ は DP マッチングの置換操作に単語間の概念距離を組み込んだ指標であり, シソーラスを利用することで求めることができる.$SEMDIST$ の値は"0〜1"となり, 値が小さいほど単語間の概念が等しいことを意味する (もし 0 なら同一概念,0 より大きい場合は 1 に近づくにつれて概念が異なっていく). これらの指標を踏まえて, 式 (5) の $dist$ の値が小さいほど用例文は入力文と類似していることを意味する (本稿では $dist$ の値が 1/4 以下となる対訳文を類似用例文として取得する). 例えば以下に示す例では $dist$ は (1+0+2*0)/(6+5)=0.090 となりこの対訳文は類似用例文として取得される (太文字:$SEMDIST$("シーディー"と"レコード"を同一概念), 下線部:挿入).

入力文:"**シーディー** を レンタル し たい です"

用例文:”**レコード** を レンタル し たい”ー”I want to rent a record”

本稿では「有効な類似用例の取得」を中心に評価するために, 類似用例文取得は以下の条件下で行なった.

類似用例文取得の条件
条件 1
用例データベースにテスト文と一致する用例文 (このような用例文は入力文の翻訳に最も効果的) がしばしば存在するが,1 文でも多くの入力文にシソーラスを適用したいため, 実験ではそのような用例文を除外した状況で類似用例文を取得する

条件 2
単語間の概念距離求める $SEMDIST$ が 1 の場合,DP マッチングの置換 S と変わりがない (すわなち, シソーラスを利用しなくても求まる) ため, この状況下での用例文は使用しない

条件 3
式 (5) で I と D のみが考慮されて求まる $dist$ はシソーラスが利用されていないので, この状況下での用例文は使用しない (すなわち, 計算される $dist$ は必ず 1 つ以上の $SEMDIST$ を含んでいる. 但し条件 2 より,$SEMDIST$=1 は考慮しない).

ここで既存のシソーラスと本稿のシソーラスから求まる $SEMDIST$ について以下に示す (但し, 両シソーラスとも”$SEMDIST$=1”は上述の条件 2 に従い, 考慮しないものとする).

既存のシソーラス
概念体系は階層を持つため, 単語間の $SEMDIST$ は”0 以上〜1 未満”となる.

本稿のシソーラス
概念体系は非階層のため, 単語間の $SEMDIST$ は”0”のみとなる.

4.1.1_b 翻訳文生成

[3] は取得した類似用例文を編集することで入力文の翻訳を得る. はじめに前節で求めた類似用例文側の $SEMDIST$ が考慮された単語の翻訳を対訳文内で獲得 (3.1 節の Step1 と同一の対訳辞書を利用) する. もし獲得に成功したら類似用例文の両文でその部分を変数化することができる, すなわち翻訳パターンが生成されること意味する. そして, 入力文側の $SEMDIST$ が考慮された単語の翻訳を先述した翻訳パターンに適用することで入力文の翻訳を得る. 前節の例を用いると類似用例文側で $SEMDIST$ が考慮された”レコード”の翻訳”record”は対訳文内に存在するので両者を変数化すると”**X** を レンタル し たい”ー”I want to rent a **X**”という翻訳パターンが生成される. この X の部分に入力文の”シーディー”の翻訳”CD”を代入すると入力文の翻訳”I want to rent a CD”が得られる.

4.1.2 翻訳評価

機械翻訳の自動評価には BLEU[9] 等がよく利用される. しかし本稿では, 前節でも示したように「有効な類似用例の取得」を中心に評価を行なうために,4.1.1_b 節の翻訳パターンを主観評価することにした. 評価概要を表 4 に示す. 翻訳パターンの決定に関して,1 つの入力文に複数の翻訳パターンが生成される場合, 各翻訳パターンの生成頻度を考慮し, 最も多く生成されるものを採用する. もし上位の生成頻度が同じ場合, 適切な翻訳パターンを手動で選択し決定する.

表 4:翻訳パターンの評価概要

評価	評価概要
○	入力文翻訳に利用可
△	入力文翻訳に条件付利用可
×	入力文翻訳に利用不可

※評価△は翻訳パターンに入力文の翻訳とは無関係の語を含んでいる場合などを意味する

4.2 シソーラス構築実験

4.2.1 実験

3 節で提案した手法で対訳コーパスからシソーラス構築を行った. 対訳コーパスは BTEC(日英対訳文:約 14 万文) を利用した. 前後の文脈情報は表 1 に示した BTEC の平均文長を考慮し, 各言語で前後 2 形態素に決定した. シソーラスの評価 (獲得される関連語の評価) は, 構築したシソーラスを 4 節の用例翻訳に適用し, 生成される翻訳パターンを主観評価 (関連性が低い関連語は不適切な翻訳パターンを生成するという観点) することで行なう. その為には,$relate\_score$ の閾値を決定する必要がある. 関連語の獲得は $relate\_score$ の閾値を操作することで振る舞いが変化し, 本稿では各閾値で得られる翻訳パターンを考慮することで閾値決定 ($relate\_score$ は値が小さいほど関連性が低いという性質を利用し, 閾値を 0.5 から徐々に上げていく) を行なった.

用例翻訳の実験環境は”用例データベース=50000 文”,”テスト文=500 文”の 3 セット (Set1,Set2,Set3. 全てのセットにおいてテスト文は独立) とした (用例データベース, テスト文は共に BTEC から作成).

4.2.2 実験結果

表 5 に各閾値で得られた関連語の数, 表 6 に各閾値で翻訳パターンが生成できた入力文の数, 表 7 に閾値間における表 6 の入力文の数の差分数とその翻訳パター

ンの評価を示す. 表 5 の結果より, 全ての閾値において獲得された関連語数は関連語候補の半分以下の割合となったが, 一定の量の関連語数が得られた. 又, 表 7 の結果より”閾値 0.5 → 1.0”,”閾値 1.0 → 1.5”はセット間で翻訳パターンの精度にばらつきがあるが,”閾値 1.5 → 2.0”は各セットで一定の精度が保たれていることが確認できた. 又, 閾値 1.5 の 3 セット分の入力文の翻訳パターン (94 文,79 文,89 文) と閾値 0.5 と 1.0 の同一の入力文の翻訳パターンとの間ではその精度に差は無く, 閾値 1.5 において一定の精度が保たれていることが確認できた. 以上の実験結果より閾値 1.5 以上の関連語は有効であると想定し, 後の実験では閾値 1.5 で議論を進めることにする.

表 5:各閾値で得られる関連語の数

relate_score の閾値	関連語の数
0.5	57276 (46.4%)
1.0	47705 (38.7%)
1.5	36597 (30.0%)
2.0	25755 (20.1%)

※括弧の数字は関連語候補全体に対する割合

表 6:各閾値で翻訳パターンが生成できた入力文の数

	relate_score の閾値			
	0.5	1.0	1.5	2.0
Set1	98	96	94	92
Set2	86	85	79	75
Set3	97	92	89	84

表 7:閾値間での翻訳パターン (評価○) の概要

	relate_score の閾値変動		
	0.5 → 1.0	1.0 → 1.5	1.5 → 2.0
Set1	0/2	2/2	1/2
Set2	1/1	4/6	4/4
Set3	4/5	2/3	5/5

※分母=閾値間での入力文の差分数
※分子=翻訳パターンが”評価○”を含む入力文数

4.3 シソーラスの性能比較実験

4.3.1 実験

前節で構築したシソーラス (*relate_score* の閾値:1.5) と既存のシソーラスを用いて, その性能比較を用例翻訳で生成される翻訳パターンを主観評価することで行なった. 既存のシソーラスには角川類語新辞典 (これを BASELINE), 構築したシソーラスには前節で構築したもの (これを PROPOSED) を用いた (それぞれの単語数は BASELINE で 51628 語,PROPOSED で 2462 語).

用例翻訳の実験環境は”用例データベース=135000 文 (135K) と 68000 文 (68K)”,”テスト文=500 文”とした (用例データベース, テスト文は, 共に BTEC から作成).

4.3.2 実験結果

翻訳パターンが生成できた入力文数とその翻訳パターンの評価を以下の 3 つに分けて評価する.

4.3.2-a 翻訳パターンが生成できた入力文数と翻訳パターン評価 (全体的な結果)

表 8 に結果を示す. 表 8 より PROPOSED はどちらの用例データベースサイズに関しても約 8 割近くが評価○ (135K:79.6%, 68K:78.5%) を得ることができた. この結果より提案手法によるシソーラス構築は一定の精度が保たれていることがわかった. しかし, 語彙サイズの大きい BASELINE に比べると翻訳カバレージは劣る結果となった.

表 8:翻訳パターンの評価概要

		翻訳パターン評価			合計
		○	△	×	(文)
135K	BASELINE	145	20	26	191
	PROPOSED	102	14	12	128
68K	BASELINE	129	18	24	171
	PROPOSED	84	15	8	107

4.3.2-b 両方のシソーラスによって翻訳パターンが生成できた入力文数と翻訳パターン評価

結果を表 9 に示す. 表 9 より, どちらの用例データベースサイズに関しても PROPOSED の精度が若干上回る結果となった.

表 9:翻訳パターンの評価概要

		翻訳パターン評価			合計
		○	△	×	(文)
135K	BASELINE	86	10	10	106
	PROPOSED	87	12	7	
68K	BASELINE	70	11	6	87
	PROPOSED	71	12	4	

次に PROPOSED による誤り改善事例を示す.

- **BASELINE が評価△かつ PROPOSE が評価○**
  例 1) 入力文”ブレスレットはありますか”
  正解文”do you have charm bracelets”
  ■ BASELINE による翻訳パターン
  ”**X** はありますか”-”do you have clip_on **X**”
  (X=イヤリング::earrings)

・$SEMDIST$ が考慮された単語
”ブレスレット (bracelets)”と”イヤリング”

■ PROPOSED による翻訳パターン
”**X** はありますか”-”do you have **X**”
(X=セーター::sweaters)
・$SEMDIST$ が考慮された単語
”ブレスレット (bracelets)”と”セーター”

・BASELINE が評価×かつ PROPOSE が評価○
例 2) 入力文”フライト番号は何番ですか”
正解文”what is your flight number”
■ BASELINE による翻訳パターン
”その **XY** は何ですか”-”what does that **X** **Y** mean”
(X=交通::traffic, Y=標識::sign)
・$SEMDIST$ が考慮された単語
”フライト (flight)”と”交通”
”番号 (number)”と”標識”

■ PROPOSED による翻訳パターン
”**X** 番号は何番ですか”-”your **X** number please”
(X=部屋::room)
・$SEMDIST$ が考慮された単語
”フライト (flight)”と”部屋”

例 1 は BASELINE 側で入力文の翻訳には関係の無い語 (”clip_on”) を含む翻訳パターンが生成された例である. これは BTEC はしばしば日本語側と英語側の単語翻訳がずれているケースが存在 (例として”鍵”-”room key”など. この場合は下線部の”room”が関係の無い語として含まれる可能性がある) し, 翻訳パターン生成の時に両言語の変数化が微妙に失敗したケースである.

例 2 は BASELINE 側で意味的に異なる翻訳パターンが生成された例であり, それぞれの単語は比較的概念が近い (”フライト”と”交通”, ”標識”と”番号”) が, 入力文の翻訳と文意が異なる用例を取得したケースである.

一方, 例 1,2 の PROPOSED 側は一般的に異なる概念 (”ブレスレット”と”セーター”, ”フライト”と”部屋”) であるが, 入力文の翻訳に適切な翻訳パターンが生成されている (例 2 は正解文とは異なるが, 言い換えを考慮すると十分適応できる範囲である). この結果より,PROPOSED のシソーラスは多言語の周辺文脈を考慮していることでより柔軟な関連語が獲得できていることがわかった.

4.3.2-c どちらか一方のシソーラスによって翻訳パターンが生成できた入力文数と翻訳パターン評価

結果を表 10 に示す. 表 10 より, 語彙サイズの大きい BASELINE の方が入力文の翻訳に必要な翻訳パターンの生成数が上回る結果となったが,BASELINE で生成不可能な翻訳パターンを PROPOSED により生成でき, その半分以上の割合で評価○を含む結果が得られた. 又,135K のように, 用例データベースサイズを 2 倍に増加させたにも関わらず PROPOSED のみで翻訳パターンが生成できているということは, 単に用例データベースサイズを増加させるだけでなく, シソーラスのようにその他の言語資源を充実させることも翻訳カバレージの寄与に繋がるという見解が得られた.

表 10:翻訳パターンの評価概要

		翻訳パターン評価			合計
		○	△	×	(文)
135K	BASELINE	59	10	16	85
	PROPOSED	15	2	5	22
68K	BASELINE	59	7	18	84
	PROPOSED	13	3	4	20

次に PROPOSED で全ての評価の事例を示す.

・PROPOSE で評価○
例 3) 入力文”今日は暖かいです”
正解文”it is warm today”
■ PROPOSED による翻訳パターン
→”**X** は暖かいです”-”it is warm **X**”
(X=外::outside)”
・$SEMDIST$ が考慮された単語
”今日 (today)”と”外”

例 4) 入力文”好きなスポーツは”
正解文”what is your favorite sport”
■ PROPOSED による翻訳パターン
→”好きな **X** は”-”what is your favorite **X**”
(X=食べ物::food)”
・$SEMDIST$ が考慮された単語
”スポーツ (sport)”と”食べ物”

・PROPOSE で評価△
例 5) 入力文”この近くに御手洗いはありますか”
正解文”is there a restroom near here”
■ PROPOSED による翻訳パターン
→”この近くに御勧めの **X** はありますか”-”is there a **X** that you can recommend near here”
(X=レストラン::restaurant)”
・$SEMDIST$ が考慮された単語
”手洗い (restroom)”と”レストラン”

・PROPOSE で評価×
例 6) 入力文”これは船便で送って下さい”
正解文”please send this by ship”
■ PROPOSED による翻訳パターン
→”**X** へは **Y** で送って下さい”-”please send them by **Y** to **X**”
(X=日本::japan, Y=航空便::airmail)”
・$SEMDIST$ が考慮された単語
”これ (this)”と”日本”
”船便 (ship)”と”航空便”

例 3,4 は一般的に異なる概念 (”今日”と”外”,”スポーツ”と”食べ物”) だが入力文の翻訳に有効な翻訳パターンが生成できた例である. これも先ほどの結果同様に多言語の周辺文脈を考慮していることが要因であると考えられる.

例 5 は式 (5) のモデルによる影響がある例である. すなわち, 例 5 の用例文に”御勧めの”-”that you can recommend”がなければ適切な翻訳パターンとして扱うことができる. 本稿のシソーラスはその全ての単語を形態素単位で取り扱っているので, 例えば”御勧めのレストラン”のように”A の B”という単語より長いコンテキスト (名詞句など) の考慮も必要と考えている.

例 6 は入力文の翻訳に不適切な翻訳パターンが生成された例である. 例 6 の場合,”これ”と”日本”の関連語が誤りの直接的な要因である. このような誤りを防ぐ一例として, 例えば式 (5) で $SEMDIST$ の計算は互いの単語のみを考慮しているが, 類似用例文側で $SEMDIST$ が考慮されている単語の周辺文脈と入力文側で $SEMDIST$ が考慮している単語を考慮し, 言語らしさをスコアとして組み込むことが挙げられる. 例 6 の場合,”これへは”という表現は日本語として不適切である. 言語らしさを計算するのに言語モデルの利用などが挙げられるが,PROPOSED のシソーラスは構築の際に関連語の前後の文脈情報を利用しており, これらを利用することで容易に求めることができる.

最後に, 本稿の実験では BASELINE と PROPOSED を別々に扱っていたが, 実験結果より PROPOSED を並行使用することで BASELINE のみで使用するよりも翻訳カバレージが増加するという見解も得られる. この見解より, 今後は PROPOSED の精度向上を目指すと共に,BASELIEN と並行活用することで翻訳カバレージの向上を目指していきたいと考えている.

5 おわりに

本稿では, 用例翻訳において既存の概念体系を持つシソーラスを用いることの問題点を掲げ, その問題を解決するために, 既存の概念体系とは異なるシソーラスを単一言語コーパスではなく多言語情報が利用できる対訳コーパスから構築することを提案した. 実験結果より, 提案手法により構築されたシソーラスは一定の精度が保たれ, 又, 用例翻訳実験では既存のシソーラスでは取得できない類似用例文を取得でき, 入力文の翻訳に適切な翻訳パターンが生成できたという結果が得られた. しかし, 翻訳カバレージは既存のシソーラスより劣る結果となった. 今後の課題として, カバレージが低かったことに関連して関連語候補生成の見直しや構築したシソーラスに階層的な要素の取り入れなどの検討が挙げられる. 又, 提案手法のシソーラスは形態素単位の単語がベースであったが, 今後は単語より長いコンテキスト (名詞句など) も考慮していきたい.

参考文献

[1] Peter Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A Statistical Approach to Machine Translation. Computational Linguistics, pp.79-85, 1990

[2] Makoto Nagao. A framework of a mechanical translation between Japanese and English by anology principle. In Elithorn, A. and Banerji, R. (eds.) Articial and Human Intelligence, pp.173-180, 1984

[3] Eiichiro Sumita. Example-based machine translation using DP-matching between word sequences. DDMT workshop of 39th ACL, pp.1-8, 2001

[4] 荒牧英治, 黒橋禎夫, 柏岡秀紀, 加藤直人. 用例ベース翻訳の確率的モデル化. 自然言語処理, Vol.13, No.3, pp.3-19, 2006

[5] 加藤直人. SDMT:用例翻訳への新しいアプローチ. 情報処理学会自然言語処理研究会, NL-170, pp.151-156, 2005

[6] Yves Lepage, Etienne Denoual. Thé Purest Ever Built EBMT System: No Variable, No Template, No Training, Examples, Just Examples, Only Examples. Workshop Example-Base Machine Translation at MT Summit X, pp.81-90.

[7] 梅村恭司他. 未踏テキスト用シソーラスの自動構築システムの開発. 平成 13 年度未踏ソフトウェア創造事業 http://www.ipa.go.jp/SPC/report/01fy-pro/explorat/

[8] Toshiyuki Takezawa, Genichiro Kikui. Collecting machine-translation-aided bilingual dialogues for corpus-based speech translation. EUROSPEECH, pp.2757-2760, 2003

[9] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. proc of the 40th ACL, pp.311-318, 2002