Vol.2012-SLP-92 No.15
2012/7/21

日本語ユーザ発話を用いた英語音声対話システム用統計的言語理解部の準教師つき学習

翠　輝久[†1]　水上悦雄[†1]　柏岡秀紀[†1]

本稿では音声対話システムの言語間移植を想定して，移植元のシステムとは異なる言語の音声言語理解部を，元のシステムの言語の言語理解部と，統計的機械翻訳を利用して構築する手法を提案する．統計的理解部の学習データとして，公開中の音声対話システムのログデータを利用することを考える．音声認識および機械翻訳に起因する誤りを含むログデータの中から，学習に有用なものを選択するために，折り返し翻訳を利用して，翻訳結果が原文の意味を保持しているかを調べる．日本語の音声対話システム AssisTra のログデータから提案手法により学習データを選択し，同システムの英語版の音声理解部を構築する実験を行い，音声言語理解の精度が改善できることを確認した．

Semi-supervised Learning of an SLU module for English Spoken Dialog System by inducting Japanese User Queries

Teruhisa Misu,[†1] Etsuo Mizukami[†1]
and Hideki Kashioka[†1]

This paper proposes a bootstrapping method of constructing a new spoken language understanding (SLU) system in a target language by utilizing statistical machine translation given an SLU module in some source language. The main challenge in this work is to induct unannotated automatic speech recognition results of user queries in the source language collected through a spoken dialog system, which is under public test. In order to select candidate expressions from among erroneous translation results stemming from problems with speech recognition and machine translation, we use back-translation results to check whether the translation result maintains the semantic meaning of the original sentence. We demonstrate that the proposed scheme can effectively prefer suitable sentences for inclusion in the training data as well as help improve the SLU module for the target language.

1. はじめに

音声対話システムへの入力発話に対して，発話の意味クラスを付与する音声言語理解タスクにおいて，アノテーション付き学習データを用いて分類器を学習すること (統計的音声言語理解) により，学習データに表れない未知の入力に対しても頑健な意味クラス推定ができることが報告されている[1)–4)]．このようなアプローチは一般に，大量の学習データを必要とするため，データ収集のコストを減らすための研究が行われている．本研究が扱う，対話システムを別の言語に移植する際においても，移植対象言語のアノテーション付きの学習データを大量に用意する必要がある．

言語間移植の際の統計的言語理解部構築のコストを減らすために，機械翻訳を利用する手法が提案されている[5)–7)]．たとえば，Servan ら[5)] は，意味クラスがアノテーションされたフランス語の MEDIA コーパスを翻訳して，イタリア語の音声言語理解部を構築し，翻訳結果を利用することの有効性を示している．Lefèvre ら[6)] は，英語のコーパスを利用してフランス語の音声言語理解部の構築を行い，複数の機械翻訳結果の利用方法の比較を行っている．本研究でも先行研究と同様に機械翻訳を利用した音声言語理解部の言語間移植を考える．

これまでの先行研究のほとんどは，移植元の言語において，大量のアノテーションされた想定発話データが利用可能であることを仮定し，これらのデータの翻訳結果を移植対象言語の音声言語理解部の学習データとして利用している (前述の先行研究においては実際に，数千程度の想定発話が利用されている．)．しかしながら，移植元の言語においてこのようなデータを用意するためには，専門家によるデータの書き起こし・アノテーション作業が必要になる．本研究ではこれに対して，移植元言語を対象言語として運用中の音声対話システムにより収集された，**アノテーションなしの音声認識結果**を利用することを考える．これは結果的に，移植元の言語の音声理解部が利用できることを仮定していることになる．この仮定は，運用中の音声対話システムの対象言語を拡大する場合を考えると，妥当であるといえる．また，Google 翻訳[⋆1]などの機械翻訳サービスが普及しつつあり，機械翻訳を利用することは容易である．

†1 情報通信研究機構
National Institute of Information and Communications Technology (NICT)
⋆1 http://translate.google.com/

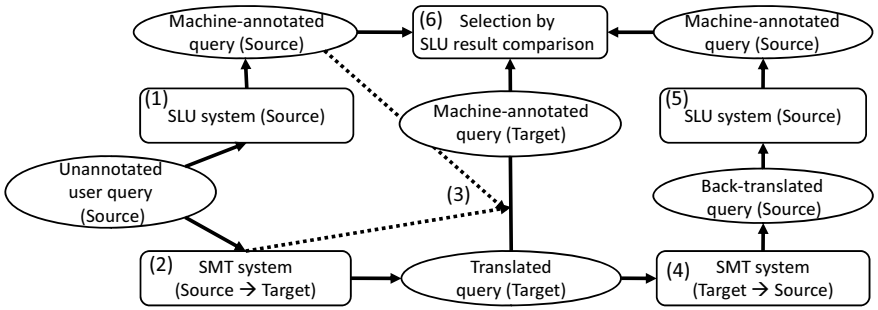

図 1 提案する文選択手法の概要

音声認識結果を用いることの主要な問題として，音声認識誤りがある．一般に，音声認識結果を用いることは，人手による書き起こしを利用する場合と比較して，学習される分類器の精度が落ちる (たとえば 9))．さらに，音声認識における誤りは，後段の機械翻訳における誤りにつながりやすい[10]．そのため，単純に音声認識結果の翻訳結果をすべて用いることが最良の方法であるとは言えない．

本研究で扱うのもう一つの問題として，言語学的に遠い言語の移植元言語のデータ (日本語) を利用して移植対象 (英語) の言語理解部を学習することが挙げられる．これまでの有効性が確認された研究では，言語学的に近い言語間の移植が行われている (たとえば，フランス語-イタリア語[5),7)]，英語-フランス語[6)])．これに対して本研究では，翻訳の過程において，文法の構造が異なるために頻繁な語順の入れ替えが起こる，言語学的に遠い言語の間での音声言語理解部の移植に取り組む．先行研究と比較して翻訳の難易度が高いため，誤りを含む翻訳結果を除去することが，より重要であると考えられる．

そこで本研究では，翻訳結果から移植対象言語の統計的言語理解部の学習データとして適切な文を選択する手法を提案する．本稿の構成は以下の通りである．2 章において，文選択を行う提案手法の概要を説明する．3 章において，統計的音声言語理解，4 章において，統計的機械翻訳モジュールについて述べる．5 章において，提案手法を用いて構築した音声言語理解部の評価を行う．6 章は結論である．

2. 音声言語理解部ポータビリティ技術

大規模コーパスに基づく機械学習手法の発達に伴い，音声認識・音声言語理解・機械翻訳の精度が向上しつつある．本研究では，これらの手法を組み合わせて，音声言語理解部を言語間移植することを目指す．すなわち，移植元の言語の音声対話システムにより収集されたユーザ発話の音声認識結果に，移植元言語の音声言語理解部によりアノテーションを行い，認識・理解結果を翻訳機を用いて移植対象言語に翻訳する．

この際に，音声認識および機械翻訳による誤りを含む認識結果を除去する必要があるため，認識結果選択の基準として，翻訳元のテキストと，テキストの折り返し翻訳結果を比較することを考える．折り返し翻訳結果と翻訳元のテキストの比較は，しばしば翻訳の正確性を確認するために利用される[11)]．翻訳の正確性を，音声言語理解に与える影響に基づいて評価するため，原文の音声言語理解結果と折り返し翻訳結果に対する音声言語理解結果を比較する (厳密には，3 章で説明する意図抽出結果の結果を比較する．)．音声言語理解結果に基づいて比較することで，翻訳に伴う単語や節の同義語への置換に頑健な文比較ができると考えられる．また，BLEU などの翻訳スコアを用いて選択する場合とは対照的に，閾値処理が必要ないことも本手法のメリットであるといえる．

提案手法による文選択の処理の流れを，図 1 に示す．また，手法は以下のように要約される．

移植元言語で運用されている音声対話システムにより収集した音声の認識結果に対して，次の処理を行う．

1. 移植元言語の音声言語理解モジュールを用いて，アノテーションを行う．
2. 音声認識結果を翻訳機を用いて移植対象言語に翻訳する．
3. 翻訳結果に対して，1. により求めたアノテーション結果を付与する．
4. 翻訳結果を折り返し翻訳する．
5. 折り返し翻訳結果に対して，移植元言語の音声言語理解モジュールを用いてアノテーションを行う．

6a. 1. と 5. の音声言語理解結果が一致した場合には，3. の結果を受理する．

6b. 一致しない場合には，3. の結果を棄却する．

本研究では，音声対話による観光案内を対象に音声言語理解部の構築を行う．移植元の言語として「日本語」を用いて「英語」の音声言語理解部を構築する．移植元言語の音声対話システムとして，我々が公開しているスマートフォンアプリケーション「AssisTra[⋆1]」のログデータを用いる．AssisTra は，京都市の観光を対象としたマルチドメイン対話システムであり，京都の観光スポット，レストラン，交通や地図の情報を提供できる．

⋆1 http://mastar.jp/assistra/index.html

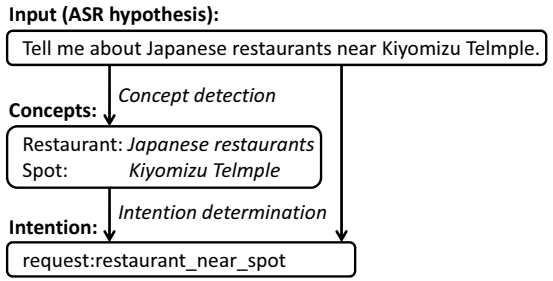

図 2 概念・発話意図アノテーションの例

3. 音声言語理解部の構成

我々が構築する音声言語理解部は，概念抽出 (固有表現抽出) 部と，発話意図抽出部から構成される．

概念抽出部は，音声認識結果から後段の対話処理部が利用するスロット情報を抽出する．すなわち，入力となる音声認識結果の単語列 $W = w_1, \ldots, w_N$ から，概念集合 $C = c_1, \ldots, c_K$ を抽出する．概念抽出部の目的は，入力単語列から，概念の集合と，それぞれの概念をを表現する部分単語列を特定することである．例として，*"Tell me about Japanese restaurants near Kiyomizu Temple"* という入力に対応する概念を図 2 に示す．この概念抽出は，しばしば BIO 記法を用いた系列データのアノテーション問題として定式化される．先ほどの例に対するアノテーション結果を以下に示す．

Tell me about Japanese restaurants near Kiyomizu Temple.

O O O B-restau. I-restau. O B-spot I-spot

概念の系列を推定するためのモデルとして，線形連鎖 CRF を採用し，予測モデルを CRF++ ツールキット[⋆1]を用いて学習した．学習に利用した素性は，単語の表層，品詞，活用形および，これらの 2-gram 情報である．観光案内タスクに対応した 20 種類の概念を定義して，アノテーションに用いる．

発話意図抽出部では，ユーザの発話意図 (これをもとに音声対話システムはユーザに対する応答を決定する) を抽出する．意図抽出部はマルチクラス SVM により学習され，学習には LIBLINEAR[⋆2] ツールキットを用いた．学習に用いる素性は，単語の表層，品詞，活用形，概念抽出部が抽出した概念，および，これらの 2-gram 情報 である．観光案内タスクに対応した 83 種類の発話意図を定義した．

4. 機 械 翻 訳

テキストの翻訳には，NICT 多言語言語翻訳研究室で開発された最新のフレーズに基づく統計的機械翻訳機 CleopATRa[⋆3][12)] を利用した．このシステムは，日英パラレルコーパス BTEC，約 70 万文から学習したモデルを利用している．このパラレルコーパスには，海外旅行を対象としたフレーズブックに出現するような，観光関連の会話が含まれている．ドメイン内の入力に対する翻訳精度は，日-英方向が 0.46，英-日方向が 0.50 である．

翻訳元の文における概念を形成する部分文字列に対応する翻訳結果の部分文字列を対応づけるために，CleopATRa の翻訳モデル (コーパスから自動獲得された翻訳用フレーズテーブル) を用いる．フレーズテーブルを用いて，最長一致の原則により概念を構成する単語の対応付けを行った．すなわち，最も多くの単語を包含する翻訳ルールから順にルールを適用する．図 3 の例では，システムは最初にフレーズ 「美味しい 日本 料理屋さん」が，フレーズテーブルに存在するかを確認し，ルールが存在しない場合には，部分文字列「日本 料理屋さん」に対応するルールを適用する．この手順を，翻訳元の概念を形成するすべての単語に対して繰り返し適用して，概念の言語間アラインメントを行う．この方法は，非常にシンプルであるが，今回のタスクにおいては，有効に機能した．これは，本タスクにおける概念の多くが名詞句であり，翻訳元の言語の特定のフレーズが，翻訳先の言語の複数のフレーズに対応することがほとんどないためであると考える．

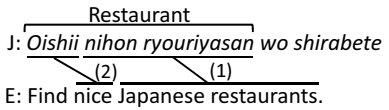

図 3 概念アラインメントの例

⋆1 http://crfpp.sourceforge.net

⋆2 http://www.csie.ntu.edu.tw/ cjlin/liblinear/

⋆3 音声翻訳システム「VoiceTra」のバックエンドシステムとして使われているものと同一である．http://mastar.jp/translation/index-en.html

表 1 学習データ・テストデータの概要

データ	発話文数	単語数
AssisTra (w/o selection)	2,950	13,007
AssisTra (with selection)	2,013	8,516
Rule	29,021	211,626
テストセット	2,537	16,095

5. 概念抽出および発話意図抽出の評価

5.1 学習・評価データ

学習データとして，移植元言語の音声対話システム「AssisTra」により収集されたユーザ発話の音声認識結果 (**AssisTra (w/o selection)**) の機械翻訳結果 2,950 文を利用する．なお，これらのデータの 10%を利用して求めた単語誤り率 (WER) は 24.8%であった．これらの発話に対する人手による書き起こしはなく，書き起こし文およびアノテーションは日本語の音声認識・音声言語理解モジュールを利用して付与し，機械翻訳により英語に翻訳したものである．これらの翻訳結果から，提案のテキスト選択手法を用いて，英語の音声言語理解部の学習に用いる文を選択した (**AssisTra (with selection)**)．提案手法によって選択された文は，2,950 文中，2,013 文であった．学習データの不足を補うために，人手により記述した文脈自由文法を用いて，データを生成した (**Rule**)．なお，文法は 871 個の生成ルールで構成されている．これらの学習データの概要を表 1 に示す．

テストセットとして,「AssisTra」を用いた被験者実験により収集した，2,537 発話の人手による翻訳結果 (**Testset**) を用いる．発話中の概念および，発話意図は人手により与えた．なお，テストセット中の上位 10 個の発話意図を持つ発話がデータ全体に占める割合は 64.7%であり，すべての発話を頻度最大のクラスに割り当てた際の正解率 (= チャンスレート) は 13.3%である．

5.2 実 験 結 果

5.2.1 参 照 手 法

学習した音声言語理解部の性能評価に先立ち，参考として，日本語の音声言語理解部 (概念抽出および発話意図抽出) の性能を評価した．これらのモジュールも英語のものと同様に，CRF と SVM を利用して，英語学習に用いた物とは別の学習データ (人手による文脈自由文法から生成した約 32,000 文 + 人手によりアノテーションされたユーザ発話約 3,000 文) を用いて学習している．日本語の音声言語理解部によるアノテーション性能を調べるために，テストセットのオリジナルデータ (英語のテストセットは，この日本語のテストセットを人手により翻訳・アノテーションしたものである．) を用いて評価した．この結果 (**Source language SLU module**) を表 2 に併記する．性能は十分に満足のいくものではないが，日本語の音声言語理解部を用いることで，人手によるアノテーションの約 90%に相当する性能で，AssisTra 発話の認識結果をアノテーションしたり，折り返し翻訳結果のチェックできることになる．

表 2 概念抽出 (F-measure, Precision, Recall) と発話意図抽出の性能比較

	概念抽出 F-measure (Precision (%), Recall (%))	発話意図抽出
Source language SLU module (reference)	87.8 (91.9, 84.1)	90.8
TestOnSource	70.3 (82.8, 61.1)	51.7
AssisTra (w/o selection)	56.6 (75.6, 44.6)	8.7
AssisTra (with selection) (proposed)	57.8 (79.6, 45.4)	50.4
Rule only	66.2 (78.1, 57.5)	56.8
Rule+AssisTra (w/o selection)	73.7 (83.9, 65.8)	43.4
Rule+AssisTra (with selection) (proposed)	75.0 (84.8, 67.3)	68.1
CVTestset only	85.3 (90.3, 80.8)	82.5
CVTestset+Rule	86.4 (88.8, 84.1)	83.6
CVTestset+Rule+AssisTra (w/o selection)	86.7 (89.2, 84.5)	76.7
CVTestset+Rule+AssisTra (with selection) (proposed)	86.5 (88.9, 84.3)	84.4

また，言語横断音声言語理解の代替手法として，TestOnSource 法[7] の評価を行った．この手法では，英語の言語理解部を構築する代わりに，機械翻訳機を用いて英語の入力文を日本語に翻訳し，日本語の音声言語理解部を用いて言語理解を行う．TestOnSource 法による結果 (**TestOnSource**) を表 2 に併記する．**Source language SLU module** との性能の差が，英-日翻訳に起因する性能劣化であると言える．

5.2.2 提案手法の評価

最初に，AssisTra コーパスのみが学習に利用できる場合の音声言語理解の性能を評価した．コーパスの機械翻訳結果を用いて，英語の音声言語理解部を構築した．学習テキストを提案手法により選択して学習した場合 **AssisTra (with selection)** ，選択しない場合の **AssisTra (w/o selection)** 性能を調べた．この結果を表 2 に併記する．性能の違いは，

表 3 人手による書き起こし結果を利用した場合との比較

AssisTra コーパス	人手による書き起こし	音声認識結果
選択なし	62.0	45.9
選択あり	62.0	62.3

発話意図抽出において顕著であった．テキスト選択を行わない場合の性能は，8.7%であり，チャンスレートの 13.3%と比較しても低いものであった．テキストの選択を行うことにより，概念抽出，発話意図理解の両方のおいて性能の改善が確認できた．

次に，移植対象言語でのシステムのプロトタイピングを想定して，学習データとして人手による文法が利用できる場合の評価を行った．この結果を表 2 に併記する．ルールのみが利用できる場合でも，AssisTra データのみを利用する場合よりも高い性能が得られた．しかしながら，提案手法により選択した AssisTra コーパスの翻訳結果を利用することで，TestOnSource 法と比較しても有意に高い性能が得られた．しかしながら，テキストの選択を行い場合には，概念抽出において改善が得られたものの，発話意図抽出の性能が大幅に低下した．これらの結果により，人手によるアノテーションが行われていないデータを利用する場合には，学習データを選択することが重要であることが確認できたとともに，提案手法により適切なテキストが選択できていることが確認できた．

次に別の利用ケースとして，ある程度の分量の人手によりアノテーションされた学習データが利用できる場合を想定して性能を評価した．この状況をシミュレートするために，テストデータを 5-fold cross-validation により分割して，評価を行った．すなわち，テストセットを 5 つに分割し，そのうち 4 つを学習データとして利用し，残りの一つを評価に用いる．この結果を表 2 に併記する．2,000 発話以上のアノテーション付きのデータが利用できる場合であっても，AssisTra コーパスを提案手法で選択して利用することにより (**CVTestset+Rule+AssisTra (with selection)**)，コーパスを利用しない場合 (**CVTestset+Rule**) と比較して改善が得られた (ただし，有意差なし)．しかしながら，コーパスを選択しない場合 (**CVTestset+Rule+AssisTra (w/o selection)**) には，前述の実験同様，性能が大幅に低下した．この結果は，多言語対応の音声対話システムをログデータの人手による書き起こし・アノテーションを行うことなく性能改善できる可能性があることを示唆している．

5.2.3 人手による書き起こしとの比較

最後に，AssisTra コーパスの一部を書き起こし，音声認識を利用して学習を行った場合との比較を行った．コーパスの 16.9%に当たる，500 発話を書き起こし，翻訳結果を提案手法を適用したところ，305 文が選択された．ルールから生成した学習データと，選択されたデータを利用して発話意図抽出器を作成して，評価を行った．この結果を表 Table 3 に示す．また，翻訳結果の選択を行わなかった場合，人手による書き起こしの代わりに音声認識結果を利用した場合の評価結果を表に併記する．選択による性能改善は得られなかったが，全体の 60%のデータの利用で，すべてのデータを利用した場合と同様の性能が得られたことが分かる．また，提案手法による選択をした場合には，音声認識結果を用いた場合であっても，人手による書き起こしを用いた場合と同等の性能が得られた．これにより，より多くのシステムのログデータが得られた場合にも，人手により書き起こしを行わなくても，性能改善ができる可能性が示唆されたと言える．

6. むすび

音声言語理解部の言語ポータビリティを行うための，準教師つき学習学習を提案した．音声認識誤りおよび機械翻訳誤りを含むコーパスから，学習データにふさわしい文を選択するために，折り返し翻訳を用いた選択基準を提案した．提案の選択手法の有効性を，既存システムとは別の言語の音声言語理解部を構築する実験を行い，人手による書き起こしを利用した場合と同等の性能を得られることを確認した．なお，提案手法は，移植元言語の音声言語理解部の性能に大きく依存するため，移植元言語の音声言語理解部の性能を上げることで，より適切な文選択を行えると期待される．さらに，音声言語理解のみではなく，提案手法を音声認識用言語モデルの言語間移植へ適用することも今後の課題である．

参考文献

1) E. Levin and R. Pieraccini: CHRONUS: the next generation , *1995 ARPA Spoken Language Systems Technical Workshop* (1995).
2) He, Y. and Young, S.: Spoken Language Understanding using the Hidden Vector State Model, *Speech Communication*, pp.262–275 (2006).
3) Wang, Y. and Acero, A.: Discriminative models for spoken language understanding, *Proc. ICSLP*, pp.1766–1769 (2006).
4) Hahn, S., Dinarelli, M., Raymond, C., Lefèvre, F., Lehnen, P., De Mori, R., Moschitti, A., Ney, H. and Riccardi, G.: Comparing Stochastic Approaches to Spoken Language Understanding in Multiple Languages, *IEEE Trans. on Speech and Audio Processing*, Vol.19, No.6, pp.1569–1583 (2011).
5) Servan, C., Camelin, N., Raymond, C., Béchet, F. and De Mori, R.: ON THE USE OF MACHINE TRANSLATION FOR SPOKEN LANGUAGE UNDERSTAND-

ING PORTABILITY, *Proc. ICASSP*, pp.5330–5333 (2010).
6) F. Lefèvre and F. Mairesse and S. Young: Cross-Lingual Spoken Language Understanding from Unaligned Data using Discriminative Classification Models and Machine Translation, *Proc. Interspeech*, pp.78–81 (2010).
7) Jabaian, B., Besacier, L. and Lefèvre, F.: COMBINATION OF STOCHASTIC UNDERSTANDING AND MACHINE TRANSLATION SYSTEMS FOR LANGUAGE PORTABILITY OF DIALOGUE SYSTEMS, *Proc. ICASSP*, pp. 5612–5615 (2011).
8) Bonneau-Maynard, H., Rosset, S., Ayache, C., Kuhn, A. and Mostefa, D.: Semantic Annotation of the French Media Dialog Corpus, *Proc. Interspeech* (2005).
9) Fabbrizio, G., Tur, G. and Hakkani-T ur, D.: Bootstrapping Spoken Dialog Systems with Data Reuse, *Proc. SIGDIAL* (2004).
10) Sarikaya, R., Zhou, B., Povey, D., Afify, M. and Gao, Y.: The Impact of ASR on the Speech-to-Speech Translation Performance, *Proc. ICASSP*, pp. 1289–1292 (2007).
11) Bach, N., Eck, M., Charoenpornsawat, P., Khler, T., Stker, S., Nguyen, T., Hsiaoa, R., Waibel, A., Vogel, S., Schultz, T. and Black, A.: The CMU TransTac 2007 Eyes-free and Hands-free Two-way Speech-to-Speech Translation System, *Proc. IWSLT* (2007).
12) Goh, C., Watanabe, T., Paul, M., Finch, A. and Sumita, E.: The NICT Translation System for IWSLT 2010, *Proc. IWSLT*, pp.139–146 (2010).