2L3-5

行動識別，スロット抽出および音声認識の統合による ノイズに頑健な命令理解

Noise Robust Instruction Understanding by Integrating of Action Identification, Slot Extraction and Speech Recognition

小堀　嵩博 *1 Takahiro Kobori　中村　友昭 *1 Tomoaki Nakamura　長井　隆行 *1 Takayuki Nagai　岩橋　直人 *2 Naoto Iwahashi　中野　幹生 *3 Mikio Nakano
船越　孝太郎 *3 Funakoshi Kotaro　金子　正秀 *1 Masahide Kaneko

*1電気通信大学 The University of Electro-Communications
*2岡山県立大学 Okayama Prefectural University
*3(株) ホンダ・リサーチ・インスティチュート・ジャパン Honda Research Institute Japan Co., Ltd

When users instruct a robot, the way of using natural language through speech is perhaps easiest method. Therefore, we propose a method that the robots can understand natural language instruction through speech. For each kind of instruction, there are a variety of natural language expressions (eg., the instruction “go to the kitchen” can also be stated as “head towards the kitchen”). Using the proposed method, the robots can understand natural language instructions in a noisy environment by integrating of action identification based on a support vector machine (SVM), extraction of nouns (slots) based on a conditional random field (CRF) that are required for robots to execute the action, co-occurrence relationship between actions and slots and speech recognition. Experimental results show that our method can achieve higher understanding accuracy in a noisy environment compared to a baseline method which used only one-best speech recognition result.

1. はじめに

一般にロボットは事前にプログラミングされた行動を実行することが想定されており，その行動を変更する際に，ロボットのプログラムを変更する必要があった．しかし，一般のユーザが必ずしもプログラムの知識があるとは限らず，特に家庭用ロボットのユーザがプログラムを変更することは困難であるといえる．ここで，もしユーザが音声を用いてロボットに命令をすることが可能であれば，容易にロボットの行動を変更できる．そこで，本稿では音声による自然な命令に対してロボットがその命令を理解する手法を提案する．ここでの自然な命令とは人が日常的に使用する言語を用いて生成される命令であり，同じ内容の命令であっても複数の言い回しが存在し，柔軟な理解が必要となる．例えば，ロボットに台所に行ってほしい場合の命令では

- 台所に行って
- 台所に行ってきて
- 台所に向かって
- ···

というように単純な命令であっても，様々な言い回しを用いて多くの命令文を生成することができる．また，ユーザが命令をロボットに発話する際には，少なからずノイズが発生する．特に家庭用ロボットにおいては人の会話などの生活音による様々なノイズが発生し，音声認識に誤認識が生じる可能性が高い．そのため，ノイズによる誤認識の影響を抑える必要があり，誤認識が多少発生したとしても言語理解を実行できるノイズに対するロバスト性が必要となる．そのような柔軟な理解とノイズに対するロバスト性は，様々なタスクが要求される家庭用ロボットにおいて重要である．

近年，ロボットによる音声命令理解への期待は高まっており，家庭用ロボットの性能を競う大会 RoboCup@Home において，音声命令の理解性能を競う課題 General Purpose Service Robot(GPSR) が行われている [RoboCup]．しかし，いまだ高得点を獲得できるチームは少なく非常に難しい課題となっている．本稿では，この GPSR タスクを対象とし，Support Vector Machine(SVM) による行動識別と Conditional Random Field(CRF) による行動に必要な名詞（スロット）抽出を組み合わせ柔軟な命令理解を行う．

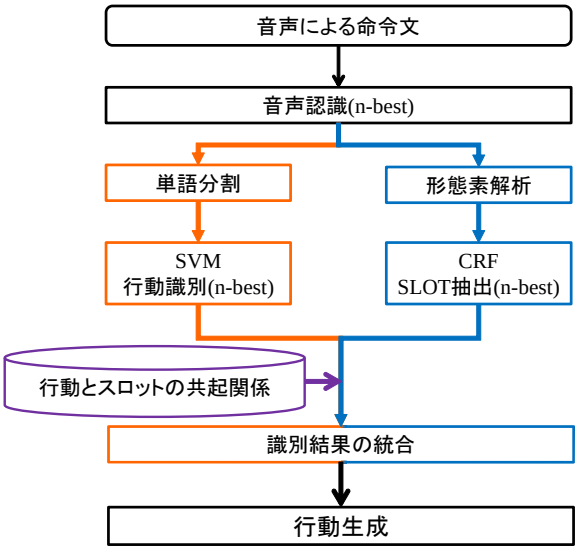

図 1: 提案手法の概要

図 1 に提案手法の概要を示す．まず，入力された音声命令の n-best の認識結果が得られ，それぞれの認識された命令文に対して形態素解析を行う．その表層系，原形，品詞等をもとに CRF によるスロット（物体名，人名等) 抽出を行う．同時に，命令文は単語分割され Bag of Words(BoW) 表現に変換し，単語と行動との関係をもとに SVM による行動識別を行

連絡先: 小堀嵩博，電気通信大学 情報理工学部 知能機械工学科，〒 182-8585，東京都調布市調布ヶ丘 1-5-1，t_kobori@radish.ee.uec.ac.jp

表 1: 行動の種類

物体を運ぶ	物体を持ってくる
物体を把持する	物体を探す
目的地へ向かう	部屋から出る
初期位置へ戻る	自己紹介する
人についていく	人を探す
人を覚える	人を認識する
人に物を手渡す	

う．行動とスロットはそれぞれ n-best の解析結果が出力されるため，行動とスロットの共起関係を利用し，適切な組み合わせを選択する．この命令理解を音声認識結果すべてで行い，最もスコアの高い解析結果の組み合わせを最終的な言語理解結果として選択する．以上のように，音声認識，行動識別，スロット抽出，行動とスロットの共起関係を統合することで，ロボットによるロバストな命令理解を行う．

関連研究として，構文解析による命令理解手法が提案されている [板谷 11][Tenorth 10][Thomas 12]．しかし，未知語が存在する場合や音声認識結果が誤っている場合に命令文を理解することが困難な場合がある．一方，本稿では汎化性能により辞書に存在しない単語や音声認識誤りが多少起こったとしても言語理解をすることが可能である．

さらに，本稿と同様に機械学習を用いた手法が提案されている [Kollar 10][Tellex 11]．しかし，[Kollar 10] では道順をロボットが理解することを目的としている．また，[Tellex 11] では [Kollar 10] を拡張し，フォークリフトロボットに適用している．この研究では主に物を動かす命令（物をつかむ，物を運ぶ等) を理解することを目的としている．よって，提案手法で対象とする GPSR タスクのような様々な種類の命令を理解することは考慮されていない．

2. 命令文理解

2.1 命令文データ

本稿では実際に GPSR タスクで使用された命令を自動生成するソフトウェア [RoboCup Soft] を使用し，主に連続した 3 つの命令からなる命令文を収集した．また，より多様な言い回しのある命令を収集するため研究室内で RoboCup@Home に参加したことのある学生にアンケートを実施し，主に 1 文のみで構成された命令文を収集した．以下の命令文が収集した例である．

- 台所に行って，ジュースを冷蔵庫から取って，田中さんに渡して
- 冷蔵庫のジュースを取って
- 冷蔵庫からジュース持って来て

このような命令文を 1569 文収集した．抽出するスロットの種類は，物体名 ([ITEM])，人の名前 ([NAME])，行動の起点となる場所 ([LOCATIONFROM])，行動の終点となる場所 ([LOCATIONTO]) の 4 種類であり，識別する行動の種類は表 1 の 13 種類である．

2.2 音声認識

ユーザから音声により入力された命令文の認識結果として n-best の認識結果 o_n とその尤度を得ることができる．しかし，音声認識結果の尤度は 1 位の認識結果が非常に高い値となり，そのまま用いると，1 位の認識結果のみを利用することとなる．しかし，音声認識では 1 位の認識結果のみが正解となるのではなく，下位の認識結果が正解となる場合もある．よって，下位の認識結果を利用できるように，音声認識スコア $S_{sr}(o_n)$ を尤度が大きいものから 1.0，0.9，0.8，... と設定することで，極端に値が小さくなることを防ぐようにした．

2.3 Support Vector Machine(SVM) による行動識別

SVM はクラス分類を行う機械学習であり，正例，負例のラベル $\boldsymbol{y} \in \{1, -1\}$ が与えられたデータ $\boldsymbol{x}$ を最も適切に正例，負例を分離できる超平面を見つけることが目的である．ここで，重みベクトル $\boldsymbol{w}$，バイアス $\boldsymbol{b}$ の 2 つのパラメータを用いて識別関数を以下のように定義する．

$$y(\boldsymbol{x}) = \boldsymbol{w}^T \cdot \boldsymbol{x} + \boldsymbol{b} \tag{1}$$

この識別関数にデータ $\boldsymbol{x}$ が与えられると，$y(\boldsymbol{x}) \geqq 0$ ならば正のクラス，$y(\boldsymbol{x}) < 0$ ならば負のクラスというように分離される．この 2 つのパラメータを学習することで，与えられた正例，負例を適切に分離できる超平面を選択する．本稿では，この 2 クラスの分類器をペアワイズ法 (pairwise method) を使用し多クラスに拡張したものを使用する．

SVM に入力するデータは語順を無視し単語同士の共起関係を示す Bag of Words (BoW) を用いることで，命令文をベクトルに変換したものを使用する．このベクトルと行動との関係を SVM により学習する．識別した行動は表 1 の 13 種類である．

2.4 Conditional Random Field(CRF) による名詞抽出

CRF は対数線形モデルを系列ラベリング問題に適用した機械学習である．系列ラベリング問題とはいくつかの要素が連なったもの (系列) にラベルを付与し，ある連続した要素を認識させたときに，それぞれの要素のラベルを予測する問題である．例として，入力された文章に対し一つ一つの形態素に分割し品詞を推定する形態素解析があげられる．本稿では，系列ラベリングをスロット抽出に適用する．まず，命令文は形態素解析により形態素に分割される．

次に，それぞれの形態素に IOB2 タグが付与される．IOB2 タグは B_*，I_*，O という 3 つのタグから構成されている．B_* がスロットの形態素の開始を示し，I_* がスロットの途中であることを示す．また O はスロット以外の形態素であることを示す．ここで，B_*，I_* における * はスロットの種類を示しており，2.1 節で述べたスロット 4 種類 ([ITEM], [LOCATIONFROM], [LOCATIONTO], [NAME]) が入る．例えば,「オレンジジュース」であれば [B_ ITEM:オレンジ]，[I_ ITEM:ジュース] というように IOB2 タグが付与される．ある命令文の形態素を $\boldsymbol{x}$ とし，スロットの IOB2 タグを $\boldsymbol{y}$ とする．この形態素 $\boldsymbol{x}$ と IOB2 タグ $\boldsymbol{y}$ の特徴は素性と呼ばれ，素性関数を $\phi(\boldsymbol{x}, y_t, y_{t+1})$ と定義する．形態素 $\boldsymbol{x}$ に対し IOB2 タグ $\boldsymbol{y}$ を割り当てる条件付き確率をそれぞれの素性の重要度を示すパラメータである重みベクトル $\boldsymbol{w}$ と素性関数を乗算することで，以下の式のように表すことができる．

$$P(\boldsymbol{y}|\boldsymbol{x}) \propto \exp(\boldsymbol{w} \cdot \phi(\boldsymbol{x}, y_t, y_{t+1})) \tag{2}$$

本稿では，t 番目の形態素の素性とし以下のものを用いた．

- 表層系のユニグラム [t -2, t+2]
- 表層系のバイグラム [t -1, t+1]
- 原形のユニグラム [t -2, t+2]
- 原形のバイグラム [t -2, t+2]
- 原形のトライグラム [t -2, t+2]
- 品詞のユニグラム [t -2, t+2]
- 品詞のバイグラム [t -2, t+2]
- 品詞のトライグラム [t -2, t+2]

- t 番目の形態素の原型と品詞の組
- y_t と y_{t-1} のバイグラム

[ ] は使用する形態素の範囲を示している．また，素性関数 $\phi(\boldsymbol{x}, y_t, y_{t+1})$ は表層形，原形，品詞に対して t 番目の形態素を中心とした上記の素性が存在すれば 1，なければ 0 となる要素を 1 列に並べたベクトルを生成する関数である．

パラメータの学習は，式 (2) で示した条件付き確率が学習データに対して最大となる重みパラメータ $\boldsymbol{w}$ を決定することであり，次式を最大化する．

$$L(\mathbf{W}) = \sum_{i}^{N} \log(y_i|x_i) \tag{3}$$

この学習した重みを用いることで，入力された文の形態素 $\bar{x}$ に対し IOB2 タグ $\bar{y}$ を条件付き確率を最大化することで求めることができる．

2.5 行動とスロットの共起関係に基づくスコア

SVM，CRF の解析結果はそれぞれ m-best，l-best の解析結果が出力され，ある命令文 o_n に対するそれぞれの解析結果 $a_m(o_n)$, $s_l(o_n)$ とそのスコア $S_{svm}(o_n)$, $S_{crf}(o_n)$ を得ることができる．この解析結果を統合するために行動とスロットの適切な組み合わせを表現した以下のようなスコア (共起スコア)$S_f(a_m(o_n), s_l(o_n))$ を導入する．

$$S_f = \frac{2 \times (Recall) \times (Precision)}{(Recall) + (Precision)} + \phi \tag{4}$$

$$Recall = \frac{(s_l(o_n) \text{ のうち行動 } a_m(o_n) \text{ で必要なスロット数})}{(\text{行動 } a_m(o_n) \text{ で必要なスロット数})} \tag{5}$$

$$Precision = \frac{(s_l(o_n) \text{ のうち行動 } a_m(o_n) \text{ で必要なスロット数})}{(s_l(o_n) \text{ のスロット数})} \tag{6}$$

$$\phi = \begin{cases} \epsilon & (R = 0 \text{ and } P = 0) \\ 0 & (\text{otherwise}) \end{cases} \tag{7}$$

この共起スコアは再現率 (Recall) と精度 (Precision) との調和平均である F 値の考え方を利用し，Recall は必要なスロットが抽出されているのかを評価し，Precision はスロットが過剰に抽出されていないかを評価する．すなわち，このスコアは行動に対して適切なスロットの組み合わせが抽出されたかを評価する．ここで，ϕ はスロットが 1 つも抽出されない場合微小な値 ϵ を加算ものであり，スコアが 0 となることを防いでいる．

2.6 識別結果の統合

最終的に，音声認識結果の n-best，SVM による行動識別結果の m-best，CRF によるスロット抽出結果の l-best のすべての組み合わせ $(n \times m \times l)$ を考慮し，それぞれのスコアを統合したスコアを計算することで，音声認識結果，行動，スロットを決定する．すなわち，次式を最大とする行動 $\hat{a}_m(\hat{o}_n)$，スロット $\hat{s}_l(\hat{o}_n)$，音声認識結果 $\hat{o}_n$ を決定する．

$$\begin{aligned} &\hat{a}_m(\hat{o}_n), \hat{s}_l(\hat{o}_n), \hat{o}_n = \\ &\underset{a_m(o_n), s_l(o_n), o_n}{\arg\max} \quad S_{svm}(a_m(o_n))^{\alpha} \; S_{crf}(s_l(o_n))^{\beta} \\ &\qquad S_f(a_m(o_n), s_l(o_n))^{\gamma} \; S_{sr}(o_n)^{\delta} \end{aligned} \tag{8}$$

ただし，α, β, γ, δ はそれぞれのスコアに対する重みパラメータである．

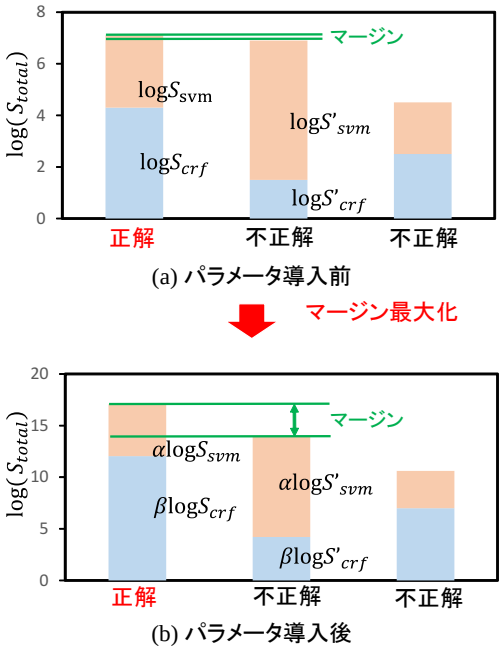

図 2: 重みパラメータの最適化

2.7 SVM による重みパラメータの推定

2.6 節の式 (8) の重みパラメータを推定するために 2 クラス分類の SVM を使用し，あいまいな結果を減らすような重みパラメータを推定する．ここでのあいまいな結果とは，正解の解析結果のスコアと不正解の解析結果で最大となるスコアとの差 (マージン) が小さいものであり，この場合，確信をもって正解と判断することができない．そのため，重みパラメータを導入し，マージンを最大化することでより確信を持って正解と不正解を判断できるようにする．例として，図 2 に重みパラメータの最適化の概要を示す．この例では，説明を簡単にするために SVM のスコア S_{svm} と CRF のスコア C_{crf} の 2 つのスコアのみを考えており，合計のスコア S_{total} を以下のように定義する．

$$S_{total} = {S_{svm}}^{\alpha} {S_{crf}}^{\beta} \tag{9}$$

ただし，乗算のままでは扱いづらいため対数を取る．

$$\log(S_{total}) = \alpha \cdot \log(S_{svm}) + \beta \cdot \log(S_{crf}) \tag{10}$$

対数を取ることで S_{svm} と S_{crf} の対数の線形和で合計のスコア S_{total} を計算することができる．ただし，$S_{svm}, S_{crf}, S'_{svm}, S'_{crf}$ は以下のように定義する．

$$S_{svm}, S_{crf} = (\text{正解となる解析結果のスコア}) \tag{11}$$

$$S'_{svm}, S'_{crf} = \underset{S'_{svm} \neq S_{svm}, S'_{crf} \neq S_{crf}}{\arg\max} (\alpha \cdot \log(S'_{svm}) + \beta \cdot \log(S'_{crf})) \tag{12}$$

図 2(a) の重みパラメータを導入する前では正解のスコアが最も高い値となっているが，中央の不正解のものとマージンが小さく確信をもって正解とは判断できない．そのため，図 2(b) のように重みパラメータを導入し，それぞれのスコアに対して重み付けをする．(b) では正解と不正解との間のスコア差が大きくなるため，より確信をもってスコアが最大の解析結果を正解と判断することができる．ここで，$\boldsymbol{x} = (\log(S_{svm}), \log(S_{crf}))$, $\boldsymbol{w} = (\alpha, \beta)$ とすると，式 (10) より全データのスコアの合計は次式となる．

$$S = \sum_{n=1}^{N} \boldsymbol{w}^T \cdot \boldsymbol{x}_n \tag{13}$$

表 2: 交差検定による命令理解精度

	認識正解文章	正答率
1-best	1443/1569	92.0%
提案手法 1	1462/1569	93.2%

マージンを最大化するために，正解のデータ $\boldsymbol{x}$ に対する $\boldsymbol{wx}$ は大きく，不正解のデータ $\bar{\boldsymbol{x}}$ に対する $\boldsymbol{w}\bar{\boldsymbol{x}}$ は小さくなるように $\boldsymbol{w}$ を決定する．ここで，正例を正解のデータと考え，負例を不正解の解析結果で最大のスコアのデータと考えることで，2 クラス分類の SVM を適用することができる．また，提案手法ではマージン差が小さいものを考えているため，$|\boldsymbol{wx}| > 1$ となるものを無視し，重み $\boldsymbol{w}$ が ∞ とならないように制約を追加する．これはソフトマージン SVM と等価である．このことから重み $\boldsymbol{w}$ は次式を最小化するものを選択する．

$$\min_{\boldsymbol{w}} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{n}^{N}(\max(0, |1 - y_n\boldsymbol{wx}_n|)) \qquad (14)$$

ここで，y_n は正例であれば 1，負例であれば -1 となる教師ラベルである．

3. 提案手法の評価実験

3.1 交差検定による評価

2.1 節で収集した命令文データを使用し 10 分割の交差検定を行った．1 回の解析で 1569 文中の 9 割を SVM, CRF の学習に使用し，残り 1 割を認識させ，正しく行動識別とスロット抽出が行えるか評価した．今回の交差検定では音声認識誤りのない命令文に対する命令の理解性能を評価するため, 式 (8) から音声認識を除いた手法 (提案手法 1) を評価した．すなわち以下の式を最大とする行動 $\hat{a}_m$ と，スロット $\hat{s}_l$ を抽出する．

$$\hat{a}_m, \hat{s}_l = \arg\max_{a_m, s_l} S_{svm}(a_m)^{\alpha}\ S_{crf}(s_l)^{\beta}\ S_f(a_m, s_l)^{\gamma} \qquad (15)$$

比較として，以下のように SVM, CRF のスコアが最大の 1-best のみを用いた手法を使用した．

$$\hat{a}_m = \arg\max_{a_m} S_{svm}(a_m) \qquad (16)$$

$$\hat{s}_l = \arg\max_{s_l} S_{crf}(s_l) \qquad (17)$$

表 2 が結果であり，1569 文中正しく認識できた割合を正答率として示している．提案手法の方がより正しく理解できた命令文が増えており，命令理解の性能を向上することができた．

3.2 音声認識による命令理解

2.1 節で集めたデータの中からランダムに 155 文を選び出し，式 (8) を用いた手法 (提案手法 2) の評価を行った．SVM, CRF に用いられる学習用データは残りの 1414 文である．比較として，3.1 で用いた音声認識スコアを除いた手法 (提案手法 1) と SVM，CRF のスコアが最大の 1-best のみを使用した手法を用いた．正答率を 155 文中で正しく認識できた割合とする．また，ノイズによる誤認識の影響を比較するため，ノイズを付与しないものと付与したものを比較した．図 3 が結果である．行動とスロットの共起関係を示すスコア $S_f(a_m(o_n), s_l(o_n))$，音声認識のスコア $S_{sr}(o_n)$ を加えていくことで，徐々に正答率を上げることが出来ている．また，ノイズが付与された環境下であっても，提案手法が最も高い正答率を示しており，音声認識誤りに対してもロバストな命令理解が可能である．

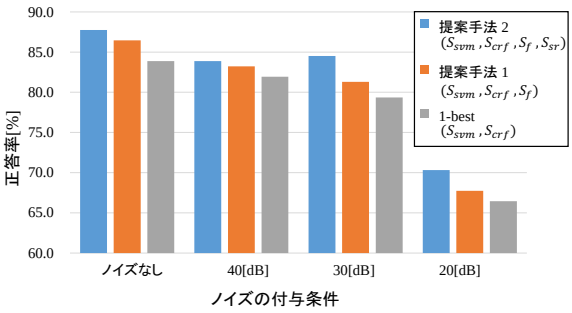

図 3: 音声認識による命令理解精度

4. まとめ

本稿では，SVM による行動識別，CRF によるスロット抽出，行動とスロットの共起関係，音声認識結果を統合した命令理解手法を提案し，RoboCup@Home での General Purpose Service Robot タスクへと適用した．SVM と CRF を単独に用いるのではなく，行動とスロットの共起関係を考え 2 つの機械学習を組み合わせた．これにより，片方の 1-best の解析結果が誤ったとしても，もう一方の解析結果から 1-best 以外の解析結果を選択できるようになり，言語理解性能を向上させることができた．また，ノイズ対策として音声認識結果を統合することで，下位の認識結果も利用できるようになり，ノイズにロバストな言語理解が可能となることを実験により示した．今後より高精度な言語理解を行うために，知覚情報との統合を図る予定である．

参考文献

[RoboCup] “RoboCup@Home,” http://www.ai.rug.nl/robocupathome/.

[板谷 11] 板谷純希, 中村友昭, 長井隆行, “ユーザとのインタラクションに基づく学習を利用したロボットのタスクプログラミング,” 第 25 回人工知能学会全国大会, 3B1-OS22c-9, 2011.

[Tenorth 10] M. Tenorth, D. Nyga, and M. Beetz “Understanding and Executing Instructions for Everyday Manipulation Tasks from the World Wide Web,” 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS), pp.1486-1491, 2010.

[Thomas 12] B. J. Thomas and O. C. Jenkins, “RoboFrameNet: Verb-Centric Semantics for Actions in Robot Middleware,” 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS), pp.4750-4755, 2012.

[Kollar 10] T. Kollar, S. Tellex, D. Roy, and N. Roy, “Toward Understanding Natural Language Directions,” Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction, pp.259-266, 2010.

[Tellex 11] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy, “Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation,” Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, pp.1507-1514, 2011.

[RoboCup Soft] “RoboCup@Home 2011 General Purpose Service Robots 文生成器（日本語版）,” http://komeisugiura.jp/software/software_jp.html.