Vol.2020-NL-246 No.9
Vol.2020-SLP-134 No.9
2020/12/2

Wisdom of Crowds を用いた音声言語理解の精度向上

吉野 幸一郎[1,2,3,a)]　池内 加奈[2]　須藤 克仁[2,3]　中村 哲[2,3]

概要：音声言語理解（SLU）とは、音声言語で与えられるユーザ発話を機械が解釈できるフレームなどの形へ変換するタスクである。既存の音声言語理解モジュールは統計的手法に基づいており、特に新しいドメインで音声言語理解モジュールを構築する場合には学習データの不足が問題であった。本論文では、この問題を二つの「Wisdom of Crowds」によって解決する方法を提案する。一つ目はクラウドソーシングによるデータ収集であり、二つ目はユーザの質問に他のユーザが回答するオンライン QA サイトである。クラウドソーシングによって新規ドメインのシードデータを収集し、このシードデータに類似するユーザ発話をオンライン QA サイトから獲得することで学習データを拡張する。この拡張データを用いて音声言語理解モジュールを構築した結果、シードデータが少量であっても音声言語理解の精度が向上することが確認された。

1. はじめに

スマートフォンやスマートスピーカーなどで動作する音声アプリケーションの研究開発が進展し、音声言語理解の重要性が増している。音声言語理解とは、ユーザからこれらのシステムに発せられた自然言語による要求から、どの機能をどういった形で呼び出すか識別するタスクである。古典的な空港乗り換え案内 [1]、レストラン案内 [2] に始まり、様々な音声言語理解タスクが研究されている [3]。

音声言語理解タスクにおいて、新しい機能の追加がされた場合の適応は重要な問題である [4]。一般に音声言語理解モジュールは機械学習を用いた統計的手法によって構築される。特に近年はニューラルネットワークを用いて理解器を構築するが、ニューラルネットワークを用いる場合、学習データの量が精度に直結する。新規機能や新規ドメインが追加された場合、異なるドメインで学習されたモデルを新規ドメインに適用する、転移学習 [5] などが代表的に用いられてきた。しかし、対象ドメインの学習データがほとんど存在しない状況で高精度の言語理解モジュールを構築することは容易ではない。これに対し近年、対象ドメインに対して作成された学習データ以外のデータから疑似的に学習データを作成するデータ拡張の手法が議論されている [6], [7]。しかし、こうした手法ではしばしば不自然な文が生成され、言語理解の精度低下の原因となる。また、ユーザの発話にはしばしば遠回しな言い方などが含まれるが、こうした言い回しを生成ベースの手法で作成することは難しい。いくつかの先行研究は、言い換え表現の生成モデルなどを用いてこうした多様な表現をカバーしようとしている [8], [9]。

これに対して、Web テキストなどから関連するテキストデータを収集する枠組みは、音声認識のための言語モデル構築などで用いられてきた [10], [11], [12], [13]。Web に存在するテキストは人間によって記述されていることが多く、疑似生成されたテキストと比較して自然で、実際にありうるテキスト候補が取得されやすい。Web に存在するテキストの大部分はデータ拡張に用いても精度低下を招くため、こうした手法では、少量のシードデータに対するパープレキシティ [14] や意味的な類似度 [15], [16] によって関連する文を選択していた。対話システムの研究においても、主に用例対話のために様々なデータ拡張手法や距離尺度が提案されている [17], [18]。

また、Web テキストのように不特定多数が記述したテキストを用いる手法は、クラウドソーシング（一つ目の「Wisdom of Crowds」）を用いる手法として一般に用いられている [19], [20]。クラウドソーシングを用いる場合、必要とする条件に適合するような学習データを低コストで収集できることが期待されるものの、機械学習に用いる大規模データセットを構築しようとする場合には依然としてコストの問題が存在する。

そこで本研究では、クラウドソーシングによって少量の

[1] 理化学研究所ロボティクスプロジェクト
RIKEN robotics project
[2] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
[3] 理化学研究所革新知能統合研究センター
RIKEN AIP
[a)] koichiro.yoshinoatriken.jp

ユーザ発話サンプルを収集し、これをシードとして Web テキストから同等の表現を獲得する手法を提案する。この際、クラウドソーシング以外の「Wisdom of Crowds」として特にオンライン QA サイトに着目した。オンライン QA サイトの質問文は良質な質問形式のユーザ発話が大量に存在し、音声対話システムのための音声認識用言語モデル構築などに有効であることが示されている [16]。そこで、オンライン QA サイトの質問文から音声言語理解の精度向上に有効なサンプルを選択するため、クラウドソーシングで収集された小規模のシードデータとの類似度に応じた選択を行った。実験の結果、オンライン QA サイトから抽出した質問文によって、音声言語理解の精度が向上することが示された。具体的には、7,600 文のシードデータから 120,000 文の拡張データを抽出することができ、音声言語理解の精度が最大で 35 ポイント向上した。

2. クラウドソーシングを用いた音声言語理解データの収集

本研究では、学習用の対話データなどが全くない状態から音声言語理解モジュールを構築する状況を想定する。まずクラウドソーシングによって少量データを構築し、これをシードとしてデータ拡張を行う。本節では、用いた音声言語理解タスクのデザインと、シードデータを作成するためのクラウドソーシングの過程について述べる。

2.1 音声言語理解のタスク設定

音声言語理解とは、音声認識やテキスト入力の結果である単語列 $x_1, x_2, ..., x_n$ から入力発話 X が与えられた場合に対話フレーム F を予測する問題である。対話フレームはドメイン、カテゴリ、クエリからなる [2]。ドメインは対話タスクの種類を示すもので、本研究では “動画”、“天気”、“ニュース”、“地図情報”、“店舗情報”、“レシピ” の 6 ドメインを設定した。カテゴリは、ドメインを細分類化したものである。例えば “動画” のドメインでは、“映画”、“ライブ映像” などの複数のドメインが設定されている。クエリは、ドメイン・カテゴリから決定されたフレームに入力されるスロットとその値である。本研究では “keyword”、“date”、“location”、“state”、“from”、“to”、“use”、“not_use” のスロットを定義した。本研究で取り扱う音声言語理解モジュールのタスクとしてはスロットに値が入る、入らないの二値について予測に限定した。なお、これらのドメイン、カテゴリ、クエリについては、Yahoo!JAPAN 社の音声アシスト[*1]で定義されている機能から選択した。

2.2 クラウドソーシングを用いたシード収集

定義された意図に対応する多様なリクエスト文のシードを収集するため、クラウドソーシングで状況を与えた上で発話パターンの収集を行った。作業ではユーザ発話を直接例示しない 3 文程度からなる状況説明を与え、クラウドワーカに自身ならどのような発話を行うかについて回答してもらった。以下に与えた教示と例を示す。

> **教示:** 3 文程度からなる日常生活のシチュエーションが提示されます。それぞれのシチュエーションに置かれた場合にあなたならどのように発言するかを入力して下さい。
> **例:** あなたは猫の動画が見たいです。あなたのお母さんがテレビのリモコンを持っています。あなたはお母さんに何とお願いしますか。

このような説明を 120 種類作成し、クラウドワーカーに対応するリクエスト発話を入力してもらった。例のようなやや曖昧な指示をすることで、様々なユーザ発話のバリエーションが収集されることを期待した。用意した意図は、ドメインごとにビデオが 24 個、天気が 19 個、ニュースが 18 個、地図が 18 個、買い物が 18 個、レシピが 19 個である。各種類に対して 100 人の作業者に依頼を行い、1 つの意図に対して 100 種類、合計 12,000 発話のユーザ発話バリエーションを収集した。

3. オンライン QA サイトを用いたデータ拡張

クラウドソーシングはある程度の量のデータを収集するのに有効であるものの、そのコストは収集データ量に線形に比例する。新しいドメインに対して機械学習を用いた音声言語理解モジュールを構築するにあたり、毎回全てのデータをクラウドソーシングで収集することは現実的ではない。そこで本研究ではもう一つの「Wisdom of Crowds」である、オンライン QA サイトの質問文を用いてデータ拡張を行う。オンライン QA サイトではユーザ同士が話し言葉調の表現を用いて会話をするため、音声アシスタントに話しかけるようなクエリに類似する質問文が存在する。特に音声対話用音声認識の言語モデル構築にはこうしたデータが有効であることが知られており [16]、音声言語理解モジュールの学習データに対しても有効であることが期待される。この際、どの質問文がどのクエリに対応するかを獲得するため、各質問文とクラウドソーシングで収集されたユーザクエリとの類似度を計算し、類似度が閾値以上のものを言語理解モジュールの拡張データとして利用する。本節ではこのデータ拡張手法について説明する。

定義された音声言語理解モジュールにおける i 種類目のユーザの意図 f_i に対して、クラウドソーシングで収集された対応する表現 $q_{i,j}(1 \leq j \leq J)$ が割り当てられている。ここで、J は意図 f_i に対してクラウドソーシングで収集されたクエリの種類数である。これらの $q_{i,j}$ に対して、オン

[*1] https://v-assist.yahoo.co.jp/, 2020 年 10 月 23 日現在

ライン QA サイトの質問文から抽出された各文 c_k との距離を計算する。$q_{i,j}$ に対して近いと判定された $\hat{c_k}$ を、f_i の意図に対応する学習データとして用いる。

文同士の類似度を計算する際、文をベクトル表現に変換して用いることが一般的である。古典的なベクトル空間モデル [21] をはじめ、発話中の単語に対する単語分散表現の平均ベクトル [22], [23]、双方向 LSTM による発話埋め込み [24], [25] など様々なベクトル作成法が試行されてきた。これに対し、近年 Transformer の構造を持つネットワークで文の周辺単語を予測する Bidirectional Encoder Representations from Transformers (BERT) [26] という手法が広く利用されるようになっている。BERT を用いる場合、文の予測モデルに対して文頭記号である [CLS] を文ベクトルと見なして利用することが一般的に行われている。

そこで本研究では、日本語 Wikipedia コーパスを用いて周辺単語予測を目的として学習した BERT モデル [27] を用いて各文に対応する [CLS] ベクトル表現を獲得した。これは、本研究で獲得したい類似する文が、単なる言い換えではなく、語用論的意味が類似する文であるためである。BERT は周辺語の予測を行うタスクに基づいて学習されているが、これは分布仮説 [28] により、類似する意味の文同士が潜在空間上で近い点に写像されていることを期待するものである。この BERT により、文 $q_{i,j}$ に対して獲得されたベクトル表現を $\mathbf{q}_{i,j}$ とする。文のベクトル表現 $\mathbf{q}_{i,j}$ と $\mathbf{c}_k$ は固定長ベクトルであるため、ベクトル同士の類似度をコサイン類似度によって求めることができる。この類似度は、

$$sim(\mathbf{q}_{i,j}, \mathbf{c}_k) = \cos(\mathbf{q}_{i,j}, \mathbf{c}_k) = \frac{\mathbf{q}_{i,j} \cdot \mathbf{c}_k}{|\mathbf{q}_{i,j}||\mathbf{c}_k|} \quad (-1 \leq sim(\cdot) \leq 1) \qquad (1)$$

として計算することができる。距離尺度は定義できたが、シードとなる各クエリ文に対してどの程度近い文を学習データに用いればよいかについては明らかでない。そこで実験では、いくつかの閾値を試行して最適な閾値を模索する。

4. 音声言語理解の実験

本研究はクラウドソーシングとオンライン QA サイトを使ったデータ拡張によって言語理解精度の向上を狙う。そこで、それぞれの拡張手法による精度向上の度合いと、データ拡張の程度による精度への影響を明らかにする。本章ではまず実験に用いた言語理解モジュールについて説明し、実験の設定と結果を述べる。

4.1 音声言語理解モジュール

2.1 節で説明したように、本研究では音声言語理解の形式としてドメイン、カテゴリ、クエリの推定を行う。このため、以前に我々が提案した LSTM に基づく音声言語理解モジュール [29][*2] を用いる。この音声言語理解モジュールは入力単語に応じて漸進的な解析を行い、DSTC2[2] のタスクにおいて高い精度を実現している。本研究ではこの漸進的言語理解モジュールを収集したタスクに合わせて学習し、音声言語理解モジュールがユーザ発話の最終単語に合わせて出力する対話状態を言語理解の結果として用いた。なお、今回定義したドメイン・カテゴリ・クエリは階層的だが、今回用いた音声言語理解モジュールではそれぞれが別々の最終層を持つため、必ずしも予測結果は連動しない。

4.2 実験設定

実験条件として、学習データにデータ拡張を行わない場合（120 種類の意図に対してそれぞれあらかじめ設定されたユーザ発話例が 1 種類）、学習データをクラウドソーシングによって拡張した場合、クラウドソーシングで拡張したデータに対してさらにオンライン QA サイトからの拡張を行った場合の比較を行った。オンライン QA サイトから得られるユーザ発話を取得するため、Yahoo!知恵袋データセット[*3]を用いた。120 の意図は 5 分割し、3 セットを学習データ、1 セットを開発データ、1 セットをテストデータとした。評価を行う開発データおよびテストデータとしては、いずれのケースもクラウドソーシングで拡張されたユーザ発話 2,000 発話（20 意図 ×100 発話、開発データ）と 2,600 発話（26 意図 ×100 発話、テストデータ）に対する精度評価を行った。この分割の際、同一の意図については同じ部分に属するように分割し、同じ意図が異なるセットに含まれないようにした。評価指標としては、対話状態推定の評価に一般に用いられる正解精度（Accuracy; 大きい方が良い）および、推定器が出力した各仮説に対する尤度と正解の one-hot 表現を比較する距離尺度（L2; 小さい方がよい）を用いる [2]。各評価スコアはドメイン、カテゴリ、クエリそれぞれに対して計算する。ただし、ドメイン、カテゴリについてはそれぞれの内容を予測する多値分類、クエリについてはクエリ値が存在する、存在しないの二値分類のスコアである。

4.3 実験結果

まず、データ拡張なし（拡張なし）、クラウドソーシングを用いた場合（+crowd）、そこからさらにオンライン QA サイトで拡張をした場合（+crowd+Web）のデータ量を表 1 に示す。$\#t$ は学習データのサンプル数であり、$\#th$ はオンライン QA サイトからデータ拡張を行う場合の類似度の閾値である。また、各手法で収集したデータに基づき構築

[*2] https://github.com/ahclab/iDST_iTTD, 2020 年 10 月 23 日現在
[*3] https://www.nii.ac.jp/dsc/idr/yahoo/chiebkr3/Y_chiebukuro.html, 2020 年 10 月 23 日現在

表 1　データ拡張の結果得られた学習データ量

	拡張なし	+crowd	+crowd+Web (th =)					
			0.90	0.88	0.86	0.84	0.82	0.80
$\#t$	74	7,400	8,048	9,868	23,312	129,093	773,874	3,592,743

表 2　各手法の精度と L2 スコア (D=ドメイン, C=カテゴリ, Q=クエリ).

評価指標	拡張手法	開発			テスト		
		D	C	Q	D	C	Q
精度: (大きい方が良い)	拡張なし	0.28	0.40	0.84	0.30	0.46	0.83
	+crowd	0.70	0.56	0.90	0.47	0.66	**0.89**
	+crowd+Web (th=0.84)	**0.89**	**0.57**	**0.92**	**0.82**	**0.72**	**0.89**
L2: (小さい方が良い)	拡張なし	0.81	0.82	0.25	0.82	0.75	0.28
	+crowd	0.47	0.72	**0.15**	0.47	0.53	**0.16**
	+crowd+Web (th=0.84)	**0.24**	**0.61**	**0.15**	**0.26**	**0.43**	**0.16**

した言語理解モジュールの評価について表 2 に示す。この結果から、ドメイン、カテゴリ、クエリのいずれについても、クラウドソーシングを使ったデータ拡張により、データ拡張を行わない場合よりも精度が向上している。また、ここからオンライン QA サイトからのデータ拡張を行った場合さらに精度が向上し、特に開発データにおけるドメインの評価では 19 ポイント、テストデータにおけるドメインの評価では 35 ポイントの大幅な改善を見せている。

オンライン QA サイトを使ってデータ拡張を行う場合、データ拡張の閾値をどうするかが問題となる。そこで閾値を変化させ、それぞれから得られた学習データから学習した言語理解モジュールの精度評価をした結果を図 1 に示す。なお、w/o はクラウドソーシングによる拡張データのみを用いる場合を示す。この結果から、開発データの傾向として閾値は th=0.84 を用いることが適当であり、これをテストデータに適用した場合ももっともよい精度が達成されることがわかる。

また、クラウドソーシングのみを利用する場合（w/o）と比較すると、いくつかのケースでは精度の向上が見られず、場合によっては悪化している。この点について、表 1 に示したデータサイズを見ると、大幅に拡張データサイズが向上する閾値 0.86 以下において、拡張なしよりもオンライン QA サイトを用いたデータ拡張のスコアが上回っている。これは、提案するオンライン QA サイトを用いたデータの全てが音声言語理解の改善に貢献するデータではないものの、閾値を緩めに取りデータ拡張を大幅にすることによって精度が改善されるということを示唆している。この定性的な分析として、いくつかクラウドおよびオンライン QA サイトからの抽出によるデータ拡張の例を表 3 に示す。この結果を見ると、クラウドソーシングはほぼ意味的に同じ意図を指す発話を収集することができている。これに対してオンライン QA サイトを拡張に用いた場合は、対話フレームは類似するものの、エンティティが異なるようなものや、近い言い回しではあるが異なる意味を持つものが拡張されている。この結果から、語用論的に同じ意図を指す様々な言い回しを収集する際にはクラウドソーシングの方が有効であることがわかるが、クラウドソーシングによってある程度の言い換え表現を獲得した後では、オンライン QA サイトから類似する言い回しを疑似的な学習データとして収集することが有効であることがわかる。また、これらの結果は Web から獲得可能な拡張データ全てが統計モデルに有効なわけではなく、適切なデータを選択するための手法が重要であるという既存研究の結論の通りである [14], [16], [30]。

5. まとめ

本研究では、クラウドソーシングとオンライン QA サイトという 2 種類の「Wisdom of Crowds」を用いて言語理解モジュールのデータ拡張を行った。オンライン QA サイトから言語理解のための拡張データを獲得するため、BERT を用いて各発話を文ベクトルへと変換して、文ベクトル同士の類似度を選択尺度として用いた。実験の結果、クラウドソーシングとオンライン QA サイトからのデータ拡張双方を利用することで、様々な言い方に頑健な音声言語理解モジュールを低コストで構築できることがわかった。

今回提案法はクラウドソーシングで収集されたデータ、つまり記述されたデータで評価したが、音声アプリケーションを目標とする場合、音声で発話されたユーザクエリに対して音声認識を適用した結果に対する評価が必要である。特に近年は音声言語理解モジュールを End-to-end で構築しようとする手法が試みられているが、こうした手法では音響特徴が重要となる。しかし、提案するような Web テキストを用いる手法では音響特徴を得ることができない。この問題を解決するための方法の一つは、マシンスピーチチェイン [31] を用いることである。マシンスピーチチェインでは、拡張されたクエリテキストから音響特徴を疑似生

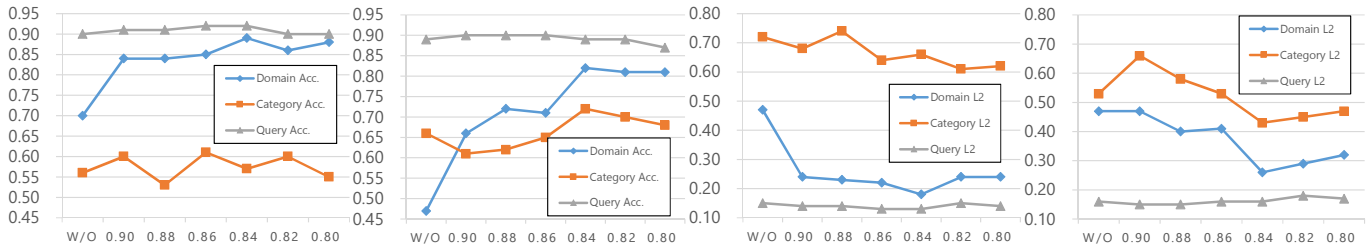

図 1 データ拡張の閾値に応じた開発、テストセットそれぞれの精度と L2 スコア。左から開発セットにおける精度、テストセットにおける精度、開発セットにおける L2、テストセットにおける L2。

表 3 各手法で拡張された学習データ例。スコアは $\mathrm{argmax}_j sim(\mathbf{q}_{i,j}, \mathbf{c}_k)$。

Added on:	文	スコア
Original	君の名は。を見たいんだけど domain=“video”, category=“movie”, keyword=“Your Name.”	-
+Crowd	君の名は。ってもう見た？	-
	君の名は。面白いんだってよ。	-
+ Web	君の名は。は見ましたか？	0.88
	君の名は。は面白かったですか？	0.87
Original	東京の天気を教えて？ domain=“weather”, category=“forecast”, location=“Tokyo”	-
+Crowd	東京って今日雨降りそう？	-
	東京の降水確率教えて	-
+ Web	明日って横浜市雨降りますか？	0.87
	今札幌で雪降ってるって本当ですか？	0.84

成することができる。また、今回は単純な距離尺度を用いたが、より発話の内容を考慮した距離尺度を検討することも必要である。例えば RoBERTa[32] のような異なる発話ベクトル変換の手法や、意味構造の利用 [16] が考えられる。

参考文献

[1] Dahl, D. A., Bates, M., Brown, M., Fisher, W., Hunicke-Smith, K., Pallett, D., Pao, C., Rudnicky, A. and Shriberg, E.: Expanding the scope of the ATIS task: The ATIS-3 corpus, *Proceedings of the workshop on Human Language Technology*, Association for Computational Linguistics, pp. 43–48 (1994).

[2] Williams, J. D., Henderson, M., Raux, A., Thomson, B., Black, A. and Ramachandran, D.: The dialog state tracking challenge series, *AI Magazine*, Vol. 35, No. 4, pp. 121–124 (2014).

[3] Hori, C., Perez, J., Higashinaka, R., Hori, T., Boureau, Y.-L., Inaba, M., Tsunomori, Y., Takahashi, T., Yoshino, K. and Kim, S.: Overview of the sixth dialog system technology challenge: Dstc6, *Computer Speech & Language*, Vol. 55, pp. 1–25 (2019).

[4] Henderson, M., Thomson, B. and Williams, J. D.: The third dialog state tracking challenge, *2014 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, pp. 324–329 (2014).

[5] Wu, C.-S., Madotto, A., Hosseini-Asl, E., Xiong, C., Socher, R. and Fung, P.: Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 808–819 (2019).

[6] Hou, Y., Liu, Y., Che, W. and Liu, T.: Sequence-to-Sequence Data Augmentation for Dialogue Language Understanding, *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1234–1245 (2018).

[7] Yoo, K. M., Shin, Y. and Lee, S.-g.: Data augmentation for spoken language understanding via joint variational generation, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 7402–7409 (2019).

[8] Saha, A., Aralikatte, R., Khapra, M. M. and Sankaranarayanan, K.: DuoRC: Towards Complex Language Understanding with Paraphrased Reading Comprehension, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1683–1693 (2018).

[9] Ray, A., Shen, Y. and Jin, H.: Robust Spoken Language Understanding via Paraphrasing, *Proceedings of Interspeech 2018*, pp. 3454–3458 (2018).

[10] Bulyko, I., Ostendorf, M. and Stolcke, A.: Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures, *Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers*, pp. 7–9 (2003).

[11] Sarikaya, R., Gravano, A. and Gao, Y.: Rapid language model development using external resources for new spoken dialog domains, *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, Vol. 1, IEEE, pp. I–573 (2005).

[12] Ng, T., Ostendorf, M., Hwang, M.-Y., Siu, M., Bulyko, I. and Lei, X.: Web-data augmented language models for mandarin conversational speech recognition, *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, Vol. 1,

IEEE, pp. I–589 (2005).
[13] Tsiartas, A., Georgiou, P. and Narayanan, S.: Language model adaptation using www documents obtained by utterance-based queries, *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, pp. 5406–5409 (2010).
[14] Misu, T. and Kawahara, T.: A bootstrapping approach for developing language model of new spoken dialogue systems by selecting web texts, *Ninth International Conference on Spoken Language Processing* (2006).
[15] Hakkani-Tur, D. and Rahim, M.: Bootstrapping language models for spoken dialog systems from the world wide web, *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Vol. 1, IEEE, pp. I–I (2006).
[16] Yoshino, K., Mori, S. and Kawahara, T.: Incorporating semantic information to selection of web texts for language model of spoken dialogue system, *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, pp. 8252–8256 (2013).
[17] Du, W. and Black, A.: Data Augmentation for Neural Online Chats Response Selection, *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, Brussels, Belgium, Association for Computational Linguistics, pp. 52–58 (2018).
[18] Henderson, M., Vulić, I., Gerz, D., Casanueva, I., Budzianowski, P., Coope, S., Spithourakis, G., Wen, T.-H., Mrkšić, N. and Su, P.-H.: Training Neural Response Selection for Task-Oriented Dialogue Systems, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 5392–5404 (2019).
[19] Zhao, L., Sukthankar, G. and Sukthankar, R.: Incremental relabeling for active learning with noisy crowdsourced annotations, *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, IEEE, pp. 728–733 (2011).
[20] Mozafari, B., Sarkar, P., Franklin, M., Jordan, M. and Madden, S.: Scaling up crowd-sourcing to very large datasets: a case for active learning, *Proceedings of the VLDB Endowment*, Vol. 8, No. 2, pp. 125–136 (2014).
[21] Salton, G., Wong, A. and Yang, C.-S.: A vector space model for automatic indexing, *Communications of the ACM*, Vol. 18, No. 11, pp. 613–620 (1975).
[22] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J.: Distributed representations of words and phrases and their compositionality, *Advances in neural information processing systems*, pp. 3111–3119 (2013).
[23] Le, Q. and Mikolov, T.: Distributed representations of sentences and documents, *International conference on machine learning*, pp. 1188–1196 (2014).
[24] Cross, J. and Huang, L.: Incremental Parsing with Minimal Features Using Bi-Directional LSTM, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 32–37 (2016).
[25] Yang, Y., Abrego, G. H., Yuan, S., Guo, M., Shen, Q., Cer, D., Sung, Y.-H., Strope, B. and Kurzweil, R.: Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax, *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, AAAI Press, pp. 5370–5378 (2019).
[26] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186 (2019).
[27] Sakata, W., Shibata, T., Tanaka, R. and Kurohashi, S.: FAQ Retrieval using Query-Question Similarity and BERT-Based Query-Answer Relevance, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1113–1116 (2019).
[28] Harris, Z. S.: Distributional structure, *Word*, Vol. 10, No. 2-3, pp. 146–162 (1954).
[29] Coman, A. C., Yoshino, K., Murase, Y., Nakamura, S. and Riccardi, G.: An Incremental Turn-Taking Model for Task-Oriented Dialog Systems, *Proceedings of Interspeech 2019*, pp. 4155–4159 (2019).
[30] Akama, R., Yokoi, S., Suzuki, J. and Inui, K.: Filtering Noisy Dialogue Corpora by Connectivity and Content Relatedness, *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2020).
[31] Tjandra, A., Sakti, S. and Nakamura, S.: Machine speech chain, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 976–989 (2020).
[32] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* (2019).