部分共有ネットワークに基づくニューラル発話意図推定の複数言語・複数タスク間での結合モデリング

増村 亮　　東中竜一郎　　政瀧 浩和　　青野裕司
日本電信電話株式会社 NTT メディアインテリジェンス研究所
{masumura.ryo, higashinaka.ryuichiro, masataki.hirokazu, aono.yushi}@lab.ntt.co.jp

1 はじめに

音声対話システムでは，対話行為 [1]，ドメイン [2]，質問種別 [3] など，複数タスクの発話意図推定を組み合わせることで，精緻な言語理解を行っている．これらの発話意図推定は，日本語や英語などの各言語の各タスクごとに，機械学習により構築されることが一般的である [4]．しかしながら，各言語の各タスクについて，学習データを十分に準備することは困難であり，言語やタスクごとに準備できるデータ量も異なる．そこで本稿では，対象の言語，対象のタスクとは異なる言語・タスクの学習データを効率的に利用することにより，発話意図推定の性能向上を狙う．

近年，発話意図推定のモデル化には，ニューラルネットワークに基づく手法 (ニューラル発話意図推定) が注目されている．ニューラル発話意図推定のネットワーク構造としては，長短期記憶を用いたリカレントニューラルネットワーク (LSTM-RNN) [5] やコンボリューショナルニューラルネットワーク [6] など様々に検討されており，素性エンジニアリングなしに高い性能を実現できることが知られている．さらに，ニューラルネットワークによるモデル化は，タスク間や言語間で知識を転移するようなモデル化にも適している．具体的には，タスクや言語に依存しない共有部分をネットワーク内に準備し，異なる言語間や異なるタスク間でモデルパラメータを共有することで，知識転移を実現できる [7–10]．発話意図推定等の文分類おいても，異なるタスク間で知識転移を行う手法が検討されており，タスク非依存のネットワークを設けることにより，各タスクの性能改善が報告されている [11–13]．また，異なる言語間で知識転移を行う手法も検討されており，言語非依存の表現を獲得することで，特に学習データの少ない言語についての性能改善が報告されている [14]．

しかしながら，既存の検討における知識転移は，異なるタスク間，または異なる言語間のどちらかに着目したものであり，言語間とタスク間の両者で同時に知識転移を行う検討はなされていない．そこで本稿では，ニューラル発話意図推定の複数言語・複数タスク間での結合モデリングを提案する．提案手法は，多対多のニューラル機械翻訳 [15–17] やマルチタスクの系列変換モデル [18] に関連する．これらのモデル化では，入力側のネットワークと出力側のネットワークを明確に区別可能である一方，ニューラル発話意図推定では，入力側と出力側のネットワークが区分けされていない．そこで我々は，ニューラル発話意図推定のネットワーク構造を，異なるタスク間で共有可能なネットワークと，異なる言語間で共有可能なネットワークに分離する．このような構造を明示的に設けることにより，複数言語，複数タスクを同時にサポートして知識転移を行うことが可能な結合モデリングを実現できる．

さらに本稿では，複数言語間，複数タスク間で質の高い知識転移を行うためのネットワーク構造を検討する．一般的に用いられる全共有ネットワークでは，タスク間で共有可能なネットワークを同一言語のデータ間で完全に共有するような構造，言語間で共有可能なネットワークを同一タスクのデータ間で完全に共有するような構造が用いられている．しかしながら，異なる言語間や異なる言語間で完全にネットワークを共有すると，一部の言語，一部のタスクの性能に劣化が生じてしまう．そこで本稿では，言語とタスクの組ごとのネットワークと共有ネットワークの両者を導入した部分共有ネットワークを提案する．これにより，異なる言語間や異なるタスク間で共有可能な情報のみが共有ネットワークに蓄積され，いずれの言語，いずれのタスクにおいても劣化なしに性能改善が期待できる．

本稿では，提案手法の有効性を評価するために，日本語と英語の 2 言語について，対話行為，拡張固有表現，質問種別の 3 種類の発話意図推定タスクのデータセットを用いる．複数言語・複数タスクで評価可能な発話意図推定や文分類のデータセットは，我々の知る限り他になく，データセットの構築も本稿の貢献である．評価実験では，複数言語・複数タスク間での結合モデリングの有効性，および部分共有ネットワークの有効性を示す．

2 ニューラル発話意図推定

発話意図推定は，入力発話の単語系列 $\mathcal{W} = \{w_1, \cdots, w_T\}$ から，ラベル $l \in \{l_1, \cdots, l_K\}$ を決定する問題である．ニューラル発話意図推定では，入力発話の単語系列が与えらた際のラベルに対する条件付き確率 $P(l|\mathcal{W}, \boldsymbol{\Theta})$ を直接ニューラルネットワークを用いてモデル化する．ここで，$\boldsymbol{\Theta}$ はモデルパラメータを表す．ニューラル発話意図推定のモデル化には，様々なモデル構造が利用できるが，本稿では双方向 LSTM-RNN(BLSTM-RNN) と Attention 機構を用いたネットワーク構造を採用する [19, 20]．以下では，モデル構造の詳細について述べる．

ニューラル発話意図推定では，最初に入力発話 $\mathcal{W}$ の各単語を連続値表現に変換する．w_t の連続値表現 $\boldsymbol{w}_t$ は (1) 式で表される．

$$\boldsymbol{w}_t = \mathtt{EMBED}(w_t; \boldsymbol{\theta}_{\mathtt{w}}) \tag{1}$$

ここで，EMBED() は単語を連続値表現に埋め込むための線形変換関数，$\boldsymbol{\theta}_{\mathtt{w}}$ はモデルパラメータを表す．次に，BLSTM-RNN を用いて各単語を前後のコンテキストを考慮した隠れ表現に変換する．t 番目の単語に

対応する隠れ表現 $\boldsymbol{h}_t$ は (2) 式で表される．

$$\boldsymbol{h}_t = \texttt{BLSTM}(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_T, t; \boldsymbol{\theta}_\texttt{h}) \tag{2}$$

ここで，BLSTM() は BLSTM-RNN の機能を持つ関数であり，$\boldsymbol{\theta}_\texttt{h}$ はモデルパラメータを表す．そして，各単語に対応する隠れ表現を用いて，発話の連続値表現を構成する．その際，Attention 機構により隠れ表現ごとの重要度を考慮した埋め込みを行う．入力発話の連続値表現 $\boldsymbol{s}$ は (3)-(4) 式に従い構成される．

$$\boldsymbol{z}_t = \texttt{tanh}(\boldsymbol{h}_t; \boldsymbol{\theta}_\texttt{z}) \tag{3}$$

$$\boldsymbol{s} = \sum_{t=1}^{T} \frac{\exp(\boldsymbol{z}_t^\top \bar{\boldsymbol{z}})}{\Sigma_{j=1}^{T} \exp(\boldsymbol{z}_j^\top \bar{\boldsymbol{z}})} \boldsymbol{h}_t \tag{4}$$

ここで，tanh() は tanh による活性化を含む非線形変換関数であり，$\boldsymbol{\theta}_\texttt{z}$ はそのモデルパラメータを表す．$\bar{\boldsymbol{z}}$ は学習可能なコンテキストベクトルであり，各隠れ表現の重要度を測る際に用いられる．最後に発話の連続表現から (5)-(6) 式に従い予測確率分布 $\boldsymbol{O}$ を算出する．

$$\boldsymbol{o} = \texttt{LINEAR}(\boldsymbol{s}; \boldsymbol{\theta}_\texttt{o}) \tag{5}$$

$$\boldsymbol{O} = \texttt{SOFTMAX}(\boldsymbol{o}) \tag{6}$$

ここで，LINEAR() は線形変換関数であり，$\boldsymbol{\theta}_\texttt{o}$ はそのモデルパラメータを表す．SOFTMAX() はソフトマックス関数を表し，$\boldsymbol{o}$ を予測確率分布に変換する．$\boldsymbol{O}$ における k 次元目の値は，$P(l_k|\mathcal{W}, \boldsymbol{\Theta})$ に対応する．なお，$\boldsymbol{\Theta}$ は $\{\boldsymbol{\theta}_\texttt{w}, \boldsymbol{\theta}_\texttt{h}, \boldsymbol{\theta}_\texttt{z}, \bar{\boldsymbol{z}}, \boldsymbol{\theta}_\texttt{o}\}$ に一致する．

モデルパラメータは，正解の確率分布と予測確率分布の交差エントロピーを最小化するように，(7) 式に従い最適化できる．

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} - \sum_{\mathcal{W} \in \mathcal{D}} \sum_{l} \hat{\boldsymbol{O}}^l_{\mathcal{W}} \log \boldsymbol{O}^l_{\mathcal{W}} \tag{7}$$

ここで，$\hat{\boldsymbol{O}}^l_{\mathcal{W}}$ と $\boldsymbol{O}^l_{\mathcal{W}}$ は，入力発話 $\mathcal{W}$ のラベル l に対する正解の確率と予測確率を表す．なお，$\mathcal{D}$ は学習データセットを表す．

3 提案手法

本節では，提案手法であるニューラル発話意図推定の複数言語・複数タスク間での結合モデルについて詳細を述べる．提案手法は，BLSTM-RNN と Attention 機構を組み合わせたニューラル発話意図推定のネットワーク構造を，異なるタスク間で共有可能なネットワークと，異なる言語間で共有可能なネットワークに明確に切り分けることで実現される．また本稿では，ネットワーク構造として，一般的な結合モデリングで採用される全共有ネットワークに加えて，部分共有ネットワークを導入する．

3.1 準備

ニューラル発話意図推定の入力側のネットワークは，単語系列であるため言語に依存するが，同一言語の異なるタスク間では共有可能である．そこで，入力発話の単語系列から BLSTM-RNN により隠れ表現の変換する部分を，タスク間で共有可能なネットワークとして，(8) 式の通り簡略化する．

$$\boldsymbol{h}_t = \texttt{W2H}(\mathcal{W}, t; \boldsymbol{\Theta}_\texttt{W2H}) \tag{8}$$

ここで，W2H() は (1)-(2) 式を含めた関数であり，$\boldsymbol{\Theta}_\texttt{W2H}$ は $\{\boldsymbol{\theta}_\texttt{w}, \boldsymbol{\theta}_\texttt{h}\}$ に一致する．

一方，ニューラル発話意図推定の出力側のネットワークは，タスクに依存するが，同一タスクの異なる言語間では共有可能である．そこで，隠れ表現系列から予測確率分布に変換する部分を，言語間で共有可能なネットワークとして，(9) 式の通り簡略化する．

$$\boldsymbol{o} = \texttt{H2O}(\boldsymbol{h}_1, \cdots, \boldsymbol{h}_T; \boldsymbol{\Theta}_\texttt{H2O}) \tag{9}$$

ここで，H2O() は (3)-(5) を含めた関数であり，$\boldsymbol{\Theta}_\texttt{H2O}$ は $\{\boldsymbol{\theta}_\texttt{z}, \bar{\boldsymbol{z}}, \boldsymbol{\theta}_\texttt{o}\}$ に一致する．

図 1 の (a) に簡略化したモデル構造によるニューラル発話意図推定を示す．このように，通常のモデル化では，単一言語，単一タスクを扱う．

3.2 全共有ネットワーク

全共有ネットワークでは，タスク間で共有可能なネットワークを同一言語のデータ間で完全に共有，言語間で共有可能なネットワークを同一タスクのデータ間で完全に共有する．全共有ネットワークのモデル構造を，図 1 の (b)-(d) に示す．グレーの部分が共有可能なネットワークであり，(b) が単一言語複数タスク，(c) が複数言語単一タスク，(d) が複数言語複数タスクのモデル構造を表している．

全共有ネットワークでは，i 番目の言語の入力発話 $\mathcal{W}^{(i)}$ における t 番目の単語を，(10) 式に従い隠れ表現に変換する．

$$\boldsymbol{h}_t = \texttt{W2H}(\mathcal{W}^{(i)}, t; \boldsymbol{\Theta}^{(i)}_\texttt{W2H}) \tag{10}$$

ここで，$\boldsymbol{\Theta}^{(i)}_\texttt{W2H}$ は i 番目の言語専用のモデルパラメータであり，i 番目の言語を扱うデータセットはいずれのタスクであってもこのパラメータを用いる．

次に，j 番目のタスクの予測確率分布 $\boldsymbol{O}^{(j)}$ の算出は (11)-(12) 式に従う．

$$\boldsymbol{o}^{(j)} = \texttt{H2O}(\boldsymbol{h}_1, \cdots, \boldsymbol{h}_T; \boldsymbol{\Theta}^{(j)}_\texttt{H2O}) \tag{11}$$

$$\boldsymbol{O}^{(j)} = \texttt{SOFTMAX}(\boldsymbol{o}^{(j)}) \tag{12}$$

ここで，$\boldsymbol{\Theta}^{(j)}_\texttt{H2O}$ は j 番目のタスク専用のモデルパラメータであり，j 番目のタスクを扱うデータセットはいずれの言語であってもこのパラメータを用いる．

3.3 部分共有ネットワーク

部分共有ネットワークでは，異なる言語間や異なるタスク間で共有するネットワークと，言語とタスクの組に対して専用のネットワークの両者を導入する．部分共有ネットワークのモデル構造を，図 1 の (e)-(g) に示す．グレーの部分が共有可能なネットワークであり，(e) が単一言語複数タスク，(f) が複数言語単一タスク，(g) が複数言語複数タスクのモデル構造を表している．

部分共有ネットワークでは，j 番目のタスクの発話意図推定を行う際の i 番目の言語の入力発話 $\mathcal{W}^{(i,j)}$ における t 番目の単語を，(13)-(15) 式に従い隠れ表現に変換する．

$$\bar{\boldsymbol{h}}^{(i,j)}_t = \texttt{W2H}(\mathcal{W}^{(i,j)}, t; \boldsymbol{\Theta}^{(i,j)}_\texttt{W2H}) \tag{13}$$

$$\bar{\boldsymbol{h}}^{(i)}_t = \texttt{W2H}(\mathcal{W}^{(i,j)}, t; \boldsymbol{\Theta}^{(i)}_\texttt{W2H}) \tag{14}$$

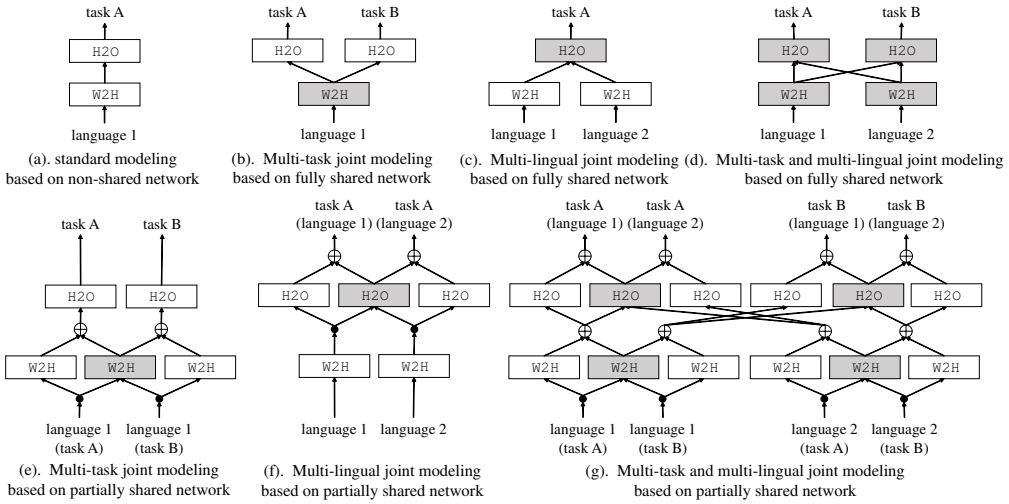

図 1: ニューラル発話意図推定における各結合モデリング手法のモデル構造.

$$\boldsymbol{h}_t^{(i,j)} = \bar{\boldsymbol{h}}_t^{(i,j)} + \bar{\boldsymbol{h}}_t^{(i)} \tag{15}$$

ここで，$\boldsymbol{\Theta}_{\texttt{W2H}}^{(i,j)}$ は i 番目の言語と j 番目のタスクの組のみが用いるモデルパラメータである．また，$\boldsymbol{\Theta}_{\texttt{W2H}}^{(i)}$ は全共有ネットワーク同様に i 番目の言語専用のモデルパラメータであり，i 番目の言語を扱うデータセットはいずれのタスクであってもこのパラメータを用いる．

次に，i 番目の言語の発話が入力された際の j 番目のタスクについての予測確率分布 $\boldsymbol{O}^{(i,j)}$ の算出は，(16)-(18) 式に従う．

$$\bar{\boldsymbol{o}}^{(i,j)} = \texttt{H2O}(\boldsymbol{h}_1^{(i,j)}, \cdots, \boldsymbol{h}_T^{(i,j)}; \boldsymbol{\Theta}_{\texttt{H2O}}^{(i,j)}) \tag{16}$$

$$\bar{\boldsymbol{o}}^{(j)} = \texttt{H2O}(\boldsymbol{h}_1^{(i,j)}, \cdots, \boldsymbol{h}_T^{(i,j)}; \boldsymbol{\Theta}_{\texttt{H2O}}^{(j)}) \tag{17}$$

$$\boldsymbol{O}^{(i,j)} = \texttt{SOFTMAX}(\bar{\boldsymbol{o}}^{(i,j)} + \bar{\boldsymbol{o}}^{(j)}) \tag{18}$$

ここで，$\boldsymbol{\Theta}_{\texttt{H2O}}^{(i,j)}$ は i 番目の言語と j 番目のタスクの組のみが用いる $\texttt{H2O}()$ のモデルパラメータである．また，$\boldsymbol{\Theta}_{\texttt{H2O}}^{(j)}$ は j 番目のタスクを扱うデータセットはいずれの言語であってもこのパラメータを用いる．

3.4 同時最適化

ニューラル発話意図推定の複数言語・複数タスク間での結合モデリングの学習は，複数言語・複数タスクで同時に最適化する．モデル全体のパラメータを $\boldsymbol{\Theta}$ と表す場合，(19) 式に従い最適化する．

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} - \sum_{\mathcal{D}^{(i,j)}} \frac{1}{|\mathcal{D}^{(i,j)}|} \sum_{\mathcal{W} \in \mathcal{D}^{(i,j)}} \sum_{l} \hat{\boldsymbol{O}}_{\mathcal{W}}^{l} \log \boldsymbol{O}_{\mathcal{W}}^{l} \tag{19}$$

ここで，$\mathcal{D}^{(i,j)}$ は i 番目の言語における j 番目のタスクの学習データセット，$|\mathcal{D}^{(i,j)}|$ は学習データセットに含まれる発話数を表す．この同時最適化をミニバッチ学習で行う場合は，各データセットの各ミニバッチを織り交ぜながら繰り返し学習することにより，徐々にモデルパラメータ $\boldsymbol{\Theta}$ を更新する．なお，データセットごとに収束のしやすさが異なるため，各データセットごとに学習率を設定し，エポックごとに各データセットの開発データを元に学習率を減少させながら学習を行う．また，学習のアーリーストッピングには，全開発データセットの平均の交差エントロピーを基準に行う．

4 評価実験

4.1 実験条件

評価実験のために，日本語と英語の 2 つの言語について，全く同一の 3 つのタスクの発話意図推定タスクを準備した．3 つのタスクは，対話行為推定，拡張固有表現推定，および質問種別推定の 3 種類であり，いずれも雑談対話システムで用いられるタスクである [4]．我々は，各データセットをそれぞれ学習データ，開発データ，テストデータに分割した．表 1 に，各データセットの発話数，およびラベル数を示す．

評価実験では，言語とタスクの組ごとに独立してモデル化を行う共有化なしのモデル化，全共有ネットワークによる結合モデリング，部分共有ネットワークによる結合モデリングの 3 手法を比較した．結合モデリングでは，複数タスク間のみでの共有，複数言語間のみでの共有，そして複数言語間・複数タスク間の同時共有を実施した．単語の連続値表現のサイズは 128，BLSTM-RNN のユニットサイズは 400，Attention 機構におけるコンテキストベクトルのサイズは 400 と，各手法で統一した．なお，学習データにおける頻度 1 以下の単語は未知語として扱った．モデルパラメータの学習では，ミニバッチ確率的勾配法により最適化を行い，開発データを用いてアーリーストッピングを行った．

4.2 実験結果

我々は各手法の評価において，モデルパラメータの初期値を変化させて 5 個のモデルを作成し，発話意図推定の平均正解率を評価した．テストセットについての正解率を表 2 に示す．(a)-(g) は図 1 に対応しており，共有化を行わない場合の結果が (a)，タスク間のみでの共有の結果が (b) と (e)，言語間のみでの共有の結果が (c) と (f)，そして言語間・タスク間の同時共有の結果が (d) と (g) である．

まず全共有ネットワークでは，共有化を行わない場合と比較して，いくつかのデータで性能改善が見られるものの，いくつかのデータでは性能劣化が確認できる．特に，言語間・タスク間の同時共有を行った場合は，共有化を行わない場合と比較して一部のデータで大きく性能劣化してしまった．これは，言語間やタ

表 2: 各評価データについての発話意図推定の正解率 (%).

		複数タスク間での結合	複数言語間での結合	日本語 対話行為	日本語 固有表現	日本語 質問種別	英語 対話行為	英語 固有表現	英語 質問種別
(a).	共有化なし	-	-	66.5	79.1	87.7	61.8	64.7	83.5
(b).	全共有	√	-	66.5	79.6	89.3	60.6	64.4	83.7
(c).		-	√	66.7	78.7	87.2	61.4	64.3	83.0
(d).		√	√	66.5	79.7	89.3	60.5	65.4	82.6
(e).	部分共有	√	-	66.6	80.9	89.4	62.0	64.8	83.7
(f).		-	√	**66.9**	79.7	88.0	61.9	65.0	83.8
(g).		√	√	**66.9**	**81.8**	**89.7**	**62.3**	**65.8**	**84.0**

表 1: 各データセットの詳細.

言語	タスク	ラベル数	学習	開発	評価
日本語	対話行為	28	201,092	4,190	4,190
日本語	固有表現	168	40,350	4,036	4,036
日本語	質問種別	15	55,328	4,257	4,257
英語	対話行為	28	25,171	3,147	3,147
英語	固有表現	168	25,005	3,230	3,230
英語	質問種別	15	22,213	2,777	2,777

スク間で共有化する場合に，完全に同一のモデルパラメータを共有して用いることが困難であることを示唆している．一方，部分共有ネットワークでは，共有化を行わない場合と比較して，性能劣化がないことが確認できる．また性能改善が得られる場合においても，全共有ネットワークよりも大きな改善効果が得られていることが分かる．特に，言語間・タスク間の同時共有により，言語間のみ，またはタスク間のみの共有化を行った場合よりも高い性能が得られてる．この結果から，部分共有ネットワークでは共有可能な情報のみ共有ネットワークに蓄積することができ，効率的に知識転移ができることが分かった．部分共有ネットワークで言語間・タスク間の同時共有を行う事により，日本語の拡張固有表現推定，日本語の質問種別推定，英語の拡張固有表現推定において，特に高い改善効果が確認できた．

5 まとめ

本稿では，音声対話システムで用いる発話意図推定の高度化を目指して，ニューラル発話意図推定の複数言語・複数タスク間での結合モデリング手法を提案した．提案手法では，BLSTM-RNN と Attention 機構を組み合わせたニューラル発話意図推定のネットワーク構造を，言語間で共有可能なネットワークと，タスク間共有可能なネットワークに明確に切り分け，異なる言語間での知識転移と異なるタスク間での知識転移の両者を可能とした．さらに，言語とタスクの組に対して専用のネットワークと共有ネットワークの両者を用いる部分共有ネットワークを提案し，データセット間で共有可能な情報のみを知識転移できる仕組みを導入した．日本語と英語の 2 言語，対話行為，拡張固有表現，質問種別の 3 タスクで設定した発話意図推定の評価実験により，部分共有ネットワークを用いた結合モデリングは，共有化を行わない場合や全共有ネットワークと比較して，効率的に性能改善できることを示した．また，言語間とタスク間の両者の知識転移を同時にサポートすることにより，言語間のみ，タスク間のみの結合モデリングと比較して，高い性能を得られることを示した．

参考文献

[1] A. Stolcke *et al.*, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[2] P. Xu and R. Sarikaya, "Contextual domain classification in spoken language understanding systems using recurrent neural network," *In Proc. ICASSP*, pp. 136–140, 2014.

[3] C.-H. Wu *et al.*, "Domain-specific FAQ retrieval using independent aspects," *ACM Transactions on Asian Language Information Processing*, vol. 4, no. 1, pp. 1–17, 2005.

[4] R. Higashinaka *et al.*, "Towards an open-domain conversational system fully based on natural language processing," *In Proc. COLING*, pp. 928–9239, 2014.

[5] S. Ravuri and A. Stolcke, "A comparative study of recurrent neural network models for lexical domain classification," *In Proc. ICASSP*, pp. 6075–6079, 2016.

[6] Y. Kim, "Convolutional neural networks for sentence classification," *In Proc. EMNLP*, pp. 1746–1751, 2014.

[7] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," *In Proc. ICML*, 2008.

[8] X. Liu *et al.*, "Representation learning using multi-task deep neural networks for semantic classification and information retrieval," *In Proc. NAACL*, pp. 912–921, 2015.

[9] Y. Liu *et al.*, "Implicit discourse relation classification via multi-task neural networks," *In Proc. AAAI*, pp. 2750–2756, 2016.

[10] X. Zhang and H. Weng, "A joint model of intent determination and slot filling for spoken language understanding," *In Proc. IJCAI*, pp. 2993–2999, 2016.

[11] P. Liu *et al.*, "Recurrent neural network for text classification with multi-task learning," *In Proc. IJCAI*, pp. 2873–2879, 2016.

[12] P. Liu *et al.*, "Deep multi-task learning with shared memory," *In Proc. EMNLP*, pp. 118–127, 2016.

[13] P. Liu *et al.*, "Adversarial multi-task learning for text classification," *In Proc. ACL*, pp. 1–10, 2017.

[14] N. Pappas and A. Popescu-Belis, "Multilingual hierarchical attention networks for document classification," *In Proc. IJCNLP*, pp. 1015–1025, 2017.

[15] O. Firat *et al.*, "Multi-way, multilingual neural machine translation with a shared attention mechanism," *In Proc. NAACL-HLT*, pp. 866–875, 2016.

[16] O. Firat *et al.*, "Multi-way, multilingual neural machine translation," *Computer Speech & Language*, pp. 236–252, 2017.

[17] H. Schwenk and M. Douze, "Learning joint multilingual sentence representations with neural machine translation," *In Proc. Workshop on Representation Learning for NLP*, pp. 157–167, 2017.

[18] M.-T. Luong *et al.*, "Multi-task sequence to sequence learning," *In Proc. ICLR*, 2016.

[19] Z. Yang *et al.*, "Hierarchical attention networks for document classification," *In Proc. NAACL-HLT*, pp. 1480–1489, 2016.

[20] P. Zhou *et al.*, "Attention-based bidirectional long short-term memory networks for relation classification," *In Proc. ACL*, pp. 207–212, 2016.