NEURAL CONFNET CLASSIFICATION: FULLY NEURAL NETWORK BASED SPOKEN UTTERANCE CLASSIFICATION USING WORD CONFUSION NETWORKS

Ryo Masumura, Yusuke Ijima, Taichi Asami, Hirokazu Masataki, Ryuichiro Higashinaka

NTT Media Intelligence Laboratories, NTT Corporation, Japan

masumura.ryo@lab.ntt.co.jp

ABSTRACT

This paper describes *neural ConfNet classification*, a novel fully neural network based spoken utterance classification method that uses word confusion networks (ConfNets). Our motivation is to establish a spoken utterance classification method that can precisely understand natural language and robustly handle automatic speech recognition (ASR) errors. Remarkable progress has been made in neural networks for accurate modeling, however, most previous methods could not handle ASR errors since they were developed for reference transcriptions. Therefore, in our work we utilized ConfNets, which are compact and efficient graph representations of ASR hypotheses. Our idea is to regard the ConfNet as a sequence of bag-of-weighted-arcs and introduce a mechanism that converts the bag-of-weighted-arcs into a continuous representation called a modified weighted sum representation. This enables us to flexibly connect ConfNets to arbitrary model structures developed for reference transcriptions. We demonstrate the effectiveness of the neural ConfNet classification in dialogue act, extended named entity, and question type classification tasks.

Index Terms— Confusion networks, neural networks, spoken utterance classification, robustness to ASR errors

1. INTRODUCTION

Spoken utterance classification tasks such as dialogue act [1], domain [2], intent [3], and question type [4] classification are essential for modern spoken dialogue systems [5]. Spoken utterance classification determines a label from an input utterance. In order to enhance the spoken utterance classification performance, both accurate modeling to understand natural languages and robust modeling of automatic speech recognition (ASR) errors are needed.

For accurate modeling, deep learning technologies have recently attracted much attention. Neural spoken utterance classification, which is a fully neural network based modeling method, demonstrates strong performance without introducing manual feature engineering. So far, various model structures such as long short-term memory recurrent neural networks (LSTM-RNNs) [6–8], convolution neural networks [9, 10], and advanced networks [11–13] have been introduced for improving classification performance.

Robust modeling of ASR errors has been also examined in various spoken language processing tasks including spoken utterance classification. It is known that classification performance deteriorates seriously due to ASR errors. Therefore, spoken utterance classification modules are often trained using ASR 1-bests rather than reference transcriptions. In addition, n-best lists [14–17], word lattices [18, 19], and word confusion networks (ConfNets) [20–23] have been utilized for taking whole ASR hypotheses into account.

However, there have been few studies on combining neural spoken utterance classification with robust modeling of ASR errors. Recently, lattice based neural spoken utterance classification methods have been proposed for addressing ASR errors [24–26]. Lattices have also been applied for neural machine translation [27, 28]. One weakness is that the model structure must be specific to lattices since lattices have complex graph structure. In other words, lattices cannot be connected to various network structures developed for reference transcriptions, i.e., simple word sequences.

This paper proposes a neural ConfNet classification method that can flexibly choose various model structures. ConfNets are more compact and efficient graph representations of ASR hypotheses than lattices. The most attractive property is that ConfNets can be represented as a linear sequential graph. Our idea is to regard a ConfNet as a sequence of bag-of-weighted-arcs and introduce a mechanism that converts the bag-of-weighted-arcs into a continuous representation. This enables our proposed method to adopt arbitrary network structures introduced in conventional neural spoken utterance classification and to compose a fully neural network based modeling using ConfNets.

For proposed neural ConfNet classification, we introduce two modeling methods to convert bag-of-weighted-arcs into a continuous representation. One is weighted sum representation, in which all word continuous representations are weighted by their posterior probability and then summed. The other is modified weighted sum representation using a self-attention mechanism [11, 12]. The representation can take the importance of words into account while considering their posterior probabilities.

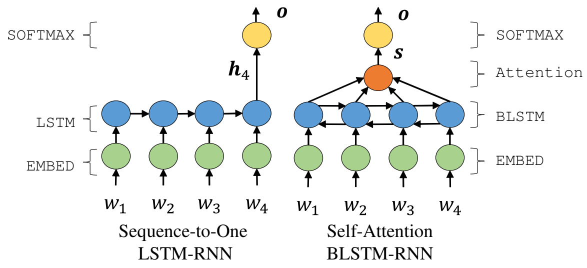

Fig. 1. Model structures of neural spoken utterance classification.

We demonstrate the effectiveness of neural ConfNet classification in three different spoken utterance classification tasks, i.e., dialogue act, extended named entity [29], and question type classification.

2. NEURAL SPOKEN UTTERANCE CLASSIFICATION

Spoken utterance classification is a problem that determines a label $l \in l_1, \cdots, l_K$ of a given utterance $\mathcal{W} = w_1, \cdots, w_T$. Neural spoken utterance classification, which is a fully neural network based modeling method, can model $P(l|\mathcal{W}, \boldsymbol{\Theta})$ in an end-to-end manner where $\boldsymbol{\Theta}$ is the model parameter.

Neural spoken utterance classification is also utilized for an n-best list based classification [26]. The n-best list includes multiple sentences generated from an ASR process. A conditional probability of a label l given an n-best list $\mathcal{L}$ is calculated as:

$$P(l|\mathcal{L}) = \sum_{\mathcal{W} \in \mathcal{L}} P(l|\mathcal{W}, \boldsymbol{\Theta}) P(\mathcal{W}), \tag{1}$$

where $P(\mathcal{W})$ denotes the posterior probability of $\mathcal{W}$ that can be calculated during an ASR process.

2.1. Modeling

Various model structures are suitable for neural spoken utterance classification. In this paper we introduce two model structures, sequence-to-one LSTM-RNN and self-attention BLSTM-RNN [11, 12]. Figure 1 shows their detailed model structures. An input is an utterance $\mathcal{W}$, and an output is predicted probabilities $\boldsymbol{o}$. The k-th dimension in $\boldsymbol{o}$ corresponds to $P(l_k|\mathcal{W}, \boldsymbol{\Theta})$.

For both model structures, each word in an input utterance $\mathcal{W}$ is first converted into a continuous representation. The continuous representation of the t-th word is defined as:

$$\boldsymbol{w}_t = \mathtt{EMBED}(w_t; \boldsymbol{\theta}_{\mathtt{w}}), \tag{2}$$

where $\mathtt{EMBED}()$ is a linear transformational function to embed a word to a continuous vector and $\boldsymbol{\theta}_{\mathtt{w}}$ is the trainable parameter.

2.1.1. *Sequence-to-One LSTM-RNN*

In sequence-to-one LSTM-RNN, each word continuous representation is converted into a hidden representation that summarizes past context information using LSTM-RNN. The hidden representation for the t-th word is calculated as:

$$\boldsymbol{h}_t = \mathtt{LSTM}(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_t; \boldsymbol{\theta}_{\mathtt{h}}), \tag{3}$$

where $\mathtt{LSTM}()$ is a function of the unidirectional LSTM-RNN layer and $\boldsymbol{\theta}_{\mathtt{h}}$ is the trainable parameter. In this case, the entire utterance information can be embedded into $\boldsymbol{h}_T$. In an output layer, predicted probabilities are produced by:

$$\boldsymbol{o} = \mathtt{SOFTMAX}(\boldsymbol{h}_T; \boldsymbol{\theta}_{\mathtt{o}}), \tag{4}$$

where $\mathtt{SOFTMAX}()$ is a transformational function with softmax activation and $\boldsymbol{\theta}_{\mathtt{o}}$ is the trainable parameter. To summarize the above, $\boldsymbol{\Theta}$ corresponds to $\{\boldsymbol{\theta}_{\mathtt{w}}, \boldsymbol{\theta}_{\mathtt{h}}, \boldsymbol{\theta}_{\mathtt{o}}\}$.

2.1.2. *Self-Attention BLSTM-RNN*

In self-attention BLSTM-RNN, each word representation is also converted into a hidden representation that takes neighboring context information into consideration. The hidden representation for the t-th word is calculated as:

$$\boldsymbol{h}_t = \mathtt{BLSTM}(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_T, t; \boldsymbol{\theta}_{\mathtt{h}}), \tag{5}$$

where $\mathtt{BLSTM}()$ is a function of the BLSTM-RNN layer. In addition, the hidden representations are summarized as a sentence representation using a self-attention mechanism that can consider the importance of individual hidden representations. The sentence continuous representation $\boldsymbol{s}$ is calculated as:

$$\boldsymbol{z}_t = \mathtt{tanh}(\boldsymbol{h}_t; \boldsymbol{\theta}_{\mathtt{z}}), \tag{6}$$

$$\boldsymbol{s} = \sum_{t=1}^{T} \frac{\exp(\boldsymbol{z}_t^\top \bar{\boldsymbol{z}})}{\Sigma_{j=1}^{T} \exp(\boldsymbol{z}_j^\top \bar{\boldsymbol{z}})} \boldsymbol{h}_t, \tag{7}$$

where $\mathtt{tanh}()$ is a non-linear transformational function with tanh activation and $\boldsymbol{\theta}_{\mathtt{z}}$ is the trainable parameter. $\bar{\boldsymbol{z}}$ is a trainable context vector, which is used for measuring the importance of individual hidden representations. In an output layer, predicted probabilities are produced by:

$$\boldsymbol{o} = \mathtt{SOFTMAX}(\boldsymbol{s}; \boldsymbol{\theta}_{\mathtt{o}}). \tag{8}$$

In this modeling, $\boldsymbol{\Theta}$ corresponds to $\{\boldsymbol{\theta}_{\mathtt{w}}, \boldsymbol{\theta}_{\mathtt{h}}, \boldsymbol{\theta}_{\mathtt{z}}, \bar{\boldsymbol{z}}, \boldsymbol{\theta}_{\mathtt{o}}\}$.

2.2. Optimization

The parameter is optimized by minimizing cross entropy loss between a reference probability and an estimated probability:

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} - \sum_{\mathcal{W} \in \mathcal{D}} \sum_{l} \hat{o}_{\mathcal{W}}^{l} \log \boldsymbol{o}_{\mathcal{W}}^{l}, \tag{9}$$

where $\hat{o}_{\mathcal{W}}^{l}$ and $\boldsymbol{o}_{\mathcal{W}}^{l}$ are respectively a reference probability and an estimated probability of label l for $\mathcal{W}$. $\mathcal{D}$ denotes the training data set.

3. NEURAL CONFNET CLASSIFICATION

This section describes our proposed neural ConfNet classification method, which is a fully neural network based modeling method using ConfNets. It can model $P(l|\mathcal{A}, \boldsymbol{\Theta})$ in an end-to-end manner where $\mathcal{A}$ is a ConfNet.

The ConfNet is a compact representation of ASR hypotheses that aligns a set of words for each position [20]. The ConfNet can be regarded as a sequence of bag-of-weighted-arcs $a_1, \cdots, a_T$. The t-th bag-of-weighted-arcs is represented as:

$$a_t = \{w_t^1, \cdots, w_t^{I_t}\}, \{P(w_t^1), \cdots, P(w_t^{I_t})\}, \qquad (10)$$

where I_t means the number of words in a_t, w_t^i is the i-th word, and $P(w_t^i)$ is a posterior probability of w_t^i. Note that a "null", which means an empty word, is also regarded as a word in the bag-of-weighted-arcs.

3.1. Modeling

In order to connect ConfNet to neural network based modeling, we convert each bag-of-weighted-arcs into a continuous representation. The continuous representation of a_t is denoted as $\boldsymbol{a}_t$. In this case, neural ConfNet classification can be structured by combining continuous representations of bag-of-weighted-arcs with model structures introduced in neural spoken utterance classification methods, i.e., sequence-to-one LSTM-RNN or self-attention BLSTM-RNN. In fact, neural ConfNet classification can be derived by replacing $\boldsymbol{w}_t$ in Eq. (3) or (5) with $\boldsymbol{a}_t$. In the following sections we introduce two modeling methods, i.e., weighted sum representation and modified weighted sum representation, for producing continuous representations of bag-of-weighted-arcs.

3.1.1. Weighted Sum Representation

The simplest of the methods is weighted sum representation, in which all word continuous representations are weighted by their posterior probability and then summed. The weighted sum representation of a_t is calculated by:

$$\boldsymbol{a}_t = \sum_{i=1}^{I_t} P(w_t^i)\mathtt{EMBED}(w_t^i; \boldsymbol{\theta}_\mathtt{w}), \qquad (11)$$

where $\mathtt{EMBED}()$ has the same function as in Eq. (2).

3.1.2. Modified Weighted Sum Representation

The posterior probability is not relevant to the importance of each word for addressing the target classification tasks. Therefore, we modify weighted sum representation by using a self-attention mechanism. The modified weighted sum representation of a_t is calculated by:

$$\boldsymbol{q}_t^i = P(w_t^i)\mathtt{EMBED}(w_t^i; \boldsymbol{\theta}_\mathtt{w}), \qquad (12)$$

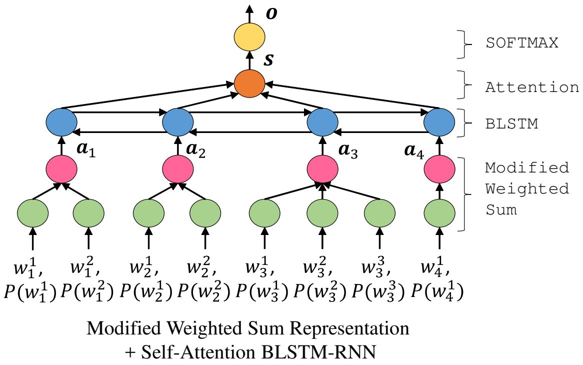

Fig. 2. Model structure of neural ConfNet classification.

$$\boldsymbol{v}_t^i = \mathtt{tanh}(\boldsymbol{q}_t^i; \boldsymbol{\theta}_\mathtt{v}), \qquad (13)$$

$$\boldsymbol{a}_t = \sum_{i=1}^{I_t} \frac{\exp(\boldsymbol{v}_t^{i\top}\bar{\boldsymbol{v}})}{\Sigma_{j=1}^{I_m}\exp(\boldsymbol{v}_t^{i\top}\bar{\boldsymbol{v}})}\boldsymbol{q}_t^i, \qquad (14)$$

where $\boldsymbol{\theta}_\mathrm{v}$ is the trainable parameter and $\bar{\boldsymbol{v}}$ is a trainable context vector, which is used for measuring the importance of individual words. Figure 2 shows a model structure that combines self-attention BLSTM-RNN with modified weighted sum representation. In this case, a model parameter $\boldsymbol{\Theta}$ means $\{\boldsymbol{\theta}_\mathtt{w}, \boldsymbol{\theta}_\mathtt{v}, \bar{\boldsymbol{v}}, \boldsymbol{\theta}_\mathrm{h}, \boldsymbol{\theta}_\mathrm{z}, \bar{\boldsymbol{z}}, \boldsymbol{\theta}_\mathrm{o}\}$.

3.2. Optimization

Neural ConfNet classification can be also optimized by minimizing cross entropy loss between a reference probability and an estimated probability:

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\mathrm{argmin}} - \sum_{\mathcal{A}\in\mathcal{D}} \sum_{l} \hat{\boldsymbol{o}}_\mathcal{A}^l \log \boldsymbol{o}_\mathcal{A}^l, \qquad (15)$$

where $\hat{\boldsymbol{o}}_\mathcal{A}^l$ and $\boldsymbol{o}_\mathcal{A}^l$ are respectively a reference probability and an estimated probability of label l for $\mathcal{A}$.

4. EXPERIMENTS

4.1. Conditions

Our experiments examined three different spoken utterance classification tasks, i.e., dialogue act (DA), extended named entity (ENE) [29, 30], and question type (QT) classification. For example, the task of ENE classification is to obtain a requested ENE type for a question. That is, for "what is the highest mountain in the world?", the ENE to be detected is "Mountain". The data sets were individually divided into training (Train), validation (Valid), and test (Test) sets. We added annotated labels to reference transcriptions. In order to investigate various ASR conditions, audio files were produced from the reference transcriptions using a homemade speech synthesizer. They were contaminated by noise for degrading

Table 2. Experimental results: utterance classification accuracy (%) for test sets.

					Sequence-to-One			Self-attention		
		Data type	Data type	Continuous representation	LSTM-RNN			BLSTM-RNN		
	Condition	in training	in testing	of bag-of-weighted-arcs	DA	ENE	QT	DA	ENE	QT
(a).	Oracle	Reference	Reference	-	63.7	65.7	82.9	63.5	65.5	82.8
(b).	Condition A	Reference	1-best	-	58.8	57.0	79.1	57.8	57.2	79.5
(c).		1-best	1-best	-	61.6	60.6	82.5	61.9	60.5	82.2
(d).		1-best	50-best	-	61.8	62.3	82.7	61.8	62.3	82.4
(e).		ConfNet	ConfNet	Weighted sum	62.7	62.7	82.7	62.9	63.0	82.7
(f).		ConfNet	ConfNet	Modified weighted sum	**63.4**	**64.0**	**82.8**	**63.6**	**65.0**	**83.0**
(g).	Condition B	Reference	1-best	-	53.2	41.8	69.1	52.0	42.0	69.3
(h).		1-best	1-best	-	58.8	49.6	79.0	58.4	50.1	78.8
(i).		1-best	50-best	-	59.8	53.2	79.6	59.3	54.2	79.0
(j).		ConfNet	ConfNet	Weighted sum	60.2	55.0	79.7	60.7	55.5	80.2
(k).		ConfNet	ConfNet	Modified weighted sum	**61.2**	**57.0**	**80.1**	**61.5**	**58.2**	**80.7**

Table 1. Experimental data sets.

	DA	ENE	QT
# of utterances in Train	26,220	26,220	26,220
# of utterances in Valid	2,622	2,622	2,622
# of utterances in Test	2,623	2,623	2,623
WER in Oracle (%)	0.0	0.0	0.0
WER Condition A (%)	11.8	14.2	14.0
WER Condition B (%)	22.9	28.9	28.1
# of labels	28	168	17
Label examples	Greeting	Age	True/false
	Apology	Company	Quantity
	Thanks	Country	Name

ASR performance. ASR hypotheses including n-bests and ConfNets were generated using a homemade speech recognizer which vocabulary size was 500K. Table 1 shows details of the data sets and word error rate (WER) in each condition. Note that utterances in each task was completely different from each other although the number of utterances was equal to each other.

For evaluation, we prepared sequence-to-one LSTM-RNNs and self-attention BLSTM-RNNs for several setups. For both modeling methods, word continuous representation size and LSTM unit size were respectively unified as 128 and 200. In training and testing for oracle condition and each ASR condition, we introduced reference transcriptions, 1-bests, 50-bests, and ConfNets. The 50-bests based testing was followed by Eq. (1). In these setups, words that appeared once or less in the training data sets were treated as unknown words. The optimization algorithm we used was Adam. The training epoch was stopped when the validation loss was not improved five consecutive times.

4.2. Results

Table 2 shows experimental results in terms of utterance classification accuracy for test sets. The proposed neural ConfNet classification is shown in (e), (f), (j), and (k). In each setup, five models were constructed by varying an initial parameter and averaged accuracy was evaluated.

The oracle condition (a), which can introduce reference transcriptions in both training and testing, demonstrated higher performance than most ASR conditions. On the other hand, reference transcription based training was not suitable for classifying 1-bests. In particular, the classification performance in (g) was substantially degraded compared to the oracle condition performance. Better classification performance was obtained for the 1-best based training than for the reference based training. Since 50-best based testing can take multiple ASR hypotheses into consideration, it provided better classification performance than the 1-best based testing. Furthermore, the proposed neural ConfNet classification, which used ConfNets in both training and testing, provided performance superior to that of n-best based neural spoken utterance classification. These results confirm that neural ConfNet classification can effectively take multiple ASR hypotheses into consideration. They also confirm that it can achieve improved performance regardless of upper network structures. In each task, the best results were achieved by (f) and (k), where modified weighted sum representation was introduced. This suggests that it is important to consider word importance in bag-of-weighted-arcs for neural ConfNet classification.

5. CONCLUSIONS

In this paper, we proposed a neural ConfNet classification that can precisely understand natural languages and robustly handle ASR errors. By introducing a mechanism that converts a sequence of bag-of-weighted-arcs to continuous representations, we enable the method to combine ConfNets with arbitrary network structures introduced in neural spoken utterance classification. Experimental results showed that the neural ConfNet classification method using modified weighted sum representation significantly outperformed n-best based neural sentence classification methods regardless of upper network structures.

6. REFERENCES

[1] Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen, "Dialogue act classification in domain-independent conversations using a deep recurrent neural network," *In Proc. International Conference on Computational Linguistics (COLING)*, pp. 2012–2021, 2016.

[2] Puyang Xu and Ruhi Sarikaya, "Contextual domain classification in spoken language understanding systems using recurrent neural network," *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 136–140, 2014.

[3] Gokhan Tur, Dilek Hakkani-Tur, Larry Heck, and Suresh Parthasarathy, "Sentence simplification for spoken language understanding," *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5628–5631, 2011.

[4] Chung-Hsien Wu, Jui-Feng Yeh, and Ming-Jun Chen, "Domain-specific FAQ retrieval using independent aspects," *ACM Transactions on Asian Language Information Processing*, vol. 4, no. 1, pp. 1–17, 2005.

[5] Ryuichiro Higashinaka, Kenji Imamura, and Toyomi Meguro, "Towards an open-domain coversational system fully based on natural language processing," *In Proc. International Conference on Computational Linguistics (COLING)*, pp. 928–9239, 2014.

[6] Suman Ravuri and Andreas Stolcke, "Recurrent neural network and LSTM models for lexical utterance classification," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 135–139, 2015.

[7] Suman Ravuri and Andreas Stolcke, "A comparative study of neural network models for lexical intent classification," *In Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 368–374, 2015.

[8] Suman Ravuri and Andreas Stolcke, "A comparative study of recurrent neural network models for lexical domain classification," *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6075–6079, 2016.

[9] Yoon Kim, "Convolutional neural networks for sentence classification," *In Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, 2014.

[10] Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu, "Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling," *In Proc. International Conference on Computational Linguistics (COLING)*, pp. 3485–3496, 2016.

[11] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu, "Attention-based bidirectional long short-term memory networks for relation classification," *In Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 207–212, 2016.

[12] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy, "Hierarchical attention networks for document classification," *In Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Tech nologies (NAACL-HLT)*, pp. 1480–1489, 2016.

[13] Naoki Sawada, Ryo Masumura, and Hiromitsu Nishizaki, "Parallel hierarchical attention networks with shared memory reader for multi-stream conversational document classification," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3311–3315, 2017.

[14] Yulan He and Steve Young, "A data-driven spoken language understanding system," *In Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 583–588, 2003.

[15] Fabrizio Morbini, Kartik Audhkhasi, Ron Artstein, Maarten Van Segbroeck, Kenji Sagae, Panayiotis Georgiou, David R. Traum, and Shri Narayanan, "A reranking approach for recognition and classification of speech input in conversational dialogue systems," *In Proc. Spoken Language Technology Workshop (SLT)*, pp. 49–54, 2012.

[16] Jean-Philippe Robichaud, Paul A. Crook, Puyang Xu, Omar Zia Khan, and Ruhi Sarikaya, "Hypotheses ranking for robust domain classification and tracking in dialogue systems," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 145–149, 2014.

[17] Omar Zia Khan, Jean-Philippe Robichaud, Paul Crook, and Ruhi Sarikaya, "Hypotheses ranking and state tracking for a multi-domain dialog system using multiple ASR alternates," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2022–2026, 2015.

[18] Murat Saraclar, "Lattice-based search for spoken utterance retrieval," *In Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 129–136, 2004.

[19] Shirin Saleem, Szu-Chen Jou, Stephan Vogel, and Tanja Schultz, "Using word lattice information for a tighter coupling in speech translation systems," *In Proc. International Conference of Spoken Language Processing (ICSLP)*, pp. 41–44, 2004.

[20] Dilek Hakkani-Tur, Frederic Bechet, Giuseppe Riccardi, and Gokhan Tur, "Beyond ASR 1-best: Using word confusion networks in spoken language understanding," *Computer Speech and Language*, vol. 20, pp. 495–514, 2006.

[21] Matthew Henderson, Milica Gasic, Blaise Thomson, Pirros Tsiakoulis, Kai Yu, and Steve Young, "Discriminative spoken language understanding using word confusion networks," *In Proc. Spoken Language Technology Workshop (SLT)*, pp. 176–181, 2012.

[22] Gokhan Tur, Anoop Deoras, and Dilek Hakkani-Tur, "Semantic parsing using word confusion networks with conditional random fields," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2579–2583, 2013.

[23] Xiaohao Yang and Jia Liu, "Using word confusion networks for slot filling in spoken language understanding," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1353–1357, 2015.

[24] Jan Svec, Adam Chylek, and Lubos Smidl, "Hierarchical discriminative model for spoken language understanding based on convolutional neural network," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1864–1868, 2015.

[25] Jan Svec, Adam Chylek, Lubos Smidl, and Pavel Ircing, "A study of different weighting schemes for spoken language understanding based on convolutional neural networks," *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6065–6069, 2016.

[26] Faisal Ladhak, Ankur Gandhe, Markus Dreyer, Lambert Mathias, Ariya Rastrow, and Bjorn Hoffmeister, "LATTICERNN: Recurrent neural networks over lattices," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 695–699, 2016.

[27] Jinsong Su, Zhixing Tan, Deyi Xiong, Rongrong Ji, Xiaodong Shi, and Yang Liu, "Lattice-based recurrent neural network encoders for neural machine translation," *In Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3302–3308, 2017.

[28] Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel, "Neural lattice-to-sequence models for uncertain inputs," *In Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1391–1400, 2017.

[29] Ryuichiro Higashinaka, Kugatsu Sadamitsu, Kuniko Saito, Toshiro Makino, and Yoshihiro Matsuo, "Creating an extended named entity dictionary from wikipedia," *In Proc. International Conference on Computational Linguistics (COLING)*, pp. 1163–1178, 2012.

[30] Satoshi Sekine and Chikashi Nobata, "Definition, dictionaries and tagger for extended named entity hierarchy," *In Proc. Language Resources and Evaluation Conference (LREC)*, pp. 1977–1980, 2004.