



Cross-Lingual Multi-Task Neural Architecture for Spoken Language Understanding

Yu-Jiang Li^{1,2}, Xuemin Zhao^{1,2}, Weiqun Xu^{1,2}, Yonghong Yan^{1,2,3}

¹Key Laboratory of Speech Acoustic and Content Understanding, Institute of Acoustics, China

²University of Chinese Academy of Sciences, China

³Xinjiang Laboratory of Minority Speech and Language Information Processing, China

yuchiang.li@outlook.com, {zhaoxuemin,xuweiqun,yanyonghong}@hcccl.ioa.ac.cn

Abstract

Cross-lingual spoken language understanding (SLU) systems traditionally require machine translation services for language portability and liberation from human supervision. However, restriction exists in parallel corpora and model architectures. Assuming reliable data are provided with human-supervision, which encourages non-parallel corpora and alleviate translation errors, this paper aims to explore cross-lingual knowledge transfer from multiple levels by taking advantage of neural architectures. We first investigate a joint model of slot filling and intent determination for SLU, which alleviates the out-of-vocabulary problem and explicitly models dependencies between output labels by combining character and word representations, bidirectional Long Short-Term Memory and conditional random fields together, while attention-based classifier is introduced for intent determination. Knowledge transfer is further operated on character-level and sequence-level, aiming to share morphological and phonological information between languages with similar alphabets by sharing character representations, and characterize the sequence with language-general and language-specific knowledge adaptively acquired by separate encoders. Experimental results on the MIT-Restaurant-Corpus and the ATIS corpora in different languages demonstrate the effectiveness of the proposed methods.

Index Terms: spoken language understanding, cross-lingual, multi-task, transfer learning

1. Introduction

Cross-lingual spoken language understanding (SLU) can be achieved in a variety of ways, especially given the increasingly available machine translation (MT) services, where no human supervision is required and language portability could be facilitated. However, this paper assumes that reliable data are already available without the restriction from MT systems. We focus on the exploration of a neural architecture that facilitates slot filling and intent determination for cross-lingual SLU.

Recent research commonly focuses on language portability, where a large of semantically-annotated utterances in the source language are available but the system is expected to be portable to a target language instead. Architectures differ from where MT systems take place, like “TrainOnSource” and “TrainOn-Target” [1]. The first one translates utterances in target languages to the source language and trains models in the source language, while the latter trains models in translated corpora instead. Though fully automatic and portable, MT-based systems may be sensitive to translation noises. Several strategies were further proposed to ameliorate such issue [2, 3, 4], e.g. the adaptive training approach that utilizes both source and trans-

lated data in the same source language for training [3, 4]. However, translated or aligned utterances may sacrifice fluency and portability to non-parallel corpora. Another potential problem comes from traditional model architectures with lack of scalability to hierarchical knowledge transfer, where semantics and pronunciation information may help.

In this work, we assume that reliable data are provided with human-supervision, aiming to reduce effect of MT errors and make it portable to non-parallel corpora. A joint neural architecture of slot filling and intent determination for SLU is first introduced, expected to be scalable in cross-lingual transfer.

Slot filling is the key component of SLU aiming to obtain semantic tags for each word in an utterance. While recurrent neural networks (RNNs) are effective for sequence modeling and have been widely adopted for slot filling [5, 6], an RNN produces a locally normalized distribution over output labels, suffers from the label bias problem similar to maximum entropy Markov models and other locally normalized models [7, 8]. To ameliorate the label bias problem, recent works [8, 9, 10, 11] incorporate dependencies between semantic labels via the conditional random fields (CRF) transition features. Another common problem that most systems with only word-level representations encountered is the out-of-vocabulary (OOV) words problem [12, 13]. Character-level representations have proven to be effective for this issue in several tasks [12, 13, 14] as morphological information that characters convey may be beneficial.

This paper investigates a general neural architecture for slot filling to solve the above two challenging problems by combining both word-level and character-level representations, bidirectional Long Short-Term Memory (BLSTM) and CRF together, similar to the model proposed for name entity recognition (NER) [14]. We further explore joint training for potential benefits from correlations between the two tasks. Similar to RNN based joint systems [9, 15, 16] trained under the framework of multi-task learning, we additionally utilize an attention layer after the BLSTM to extract words that are significant to the semantics of an utterance, which is inspired from the encoder-decoder framework [15, 16, 17].

In this paper, we propose several approaches to leverage hierarchical knowledge in both target and source languages with similar alphabets, which are not restricted to parallel corpora. Inspired by recent research on transfer learning [18], we seek to learn shared character representations between related languages where morphological and phonological knowledge may exist in common. Approaches for language-adaptive training traditionally fine-tuned on combined datasets [3, 4]. We seek to model language-general and language-specific representations explicitly, inspired from [19], language-general knowledge on the sequence level are encouraged to be compatible with differ-

ent languages via a shared BLSTM, then jointly characterize a sequence with language-specific knowledge acquired by a specific BLSTM. Gate connection is further proposed to adaptively characterize invariant and variant information of a language.

The remainder of the paper is organized as follows. Section 2 sets the baseline CharNN-BLSTM-CRF architecture for slot filling task. In section 3, we extend this architecture to jointly model intent determination and slot filling. Approaches for cross-lingual SLU will also be presented in detail. The next section details our datasets, implementation and experiments. Finally, conclusions are drawn in section 5.

2. CharNN-BLSTM-CRF architecture

2.1. Character-level representation

SLU systems traditionally require word representations, regardless of the OOV problem and noises introduced by automatic speech recognition (ASR) [20, 21]. Previous studies [12, 13, 14] have explored character representations to alleviate such problems as morphological knowledge can be obtained from characters, like the prefix or suffix of a word. Another potential benefit is the robustness to ASR errors, though a sequence of characters may not be the object word, they may convey phonological information as most ASR systems model phonemes even characters directly in acoustic models [20, 22], where correspondences exist between phonemes and graphemes to a large extent.

While character representations can be modeled by RNNs e.g. BLSTM [12], recent approaches [13, 14] have demonstrated the effectiveness of CNNs to extract morphological information from characters of a word and encode it into neural representations. We adopt the neural architecture proposed in [11], where initial character embeddings with dropout are fed into the convolution layer and then a max-pooling layer.

2.2. BLSTM

BLSTM is utilized to model the entire sequence by capturing both past and future information, which has proven to be effective in previous works [12, 15]. Given a sequence $x = (x_1, x_2, \dots, x_T)$, each input at time t is represented as a d -dimensional vector x_t , an LSTM generates a hidden state \vec{h}_t of the forward context at time t and another LSTM calculates the backward hidden state \overleftarrow{h}_t . Final representation of the hidden state at each time step is obtained by concatenating both forward and backward ones $h_t = [\vec{h}_t, \overleftarrow{h}_t]$. Stacked BLSTM can also be applied, while only one layer is used in this work.

2.3. CRFs

For sequence tagging tasks, it is beneficial to explicitly model the dependencies between labels in neighborhood and jointly decode the optimal labels. Therefore, instead of modeling tagging decisions independently as in locally normalized models, e.g. RNN, we model them jointly using CRFs [23, 24]. Specifically, linear-chain CRF [24] is adopted for efficiency and convenience in this work. Given an input with a sequence of vectors $x = (x_1, x_2, \dots, x_T)$ and corresponding labels $y = (y_1, y_2, \dots, y_T)$, the global feature vector has the form :

$$F(y, x) = \sum_{t,i} \lambda_i t_i(y_{t-1}, y_t, x, t) + \sum_{t,j} \mu_j s_j(y_t, x, t) \quad (1)$$

where $t_i(y_{t-1}, y_t, x, t)$ is a transition feature function of the entire observation sequence and labels at time step $t - 1$ and

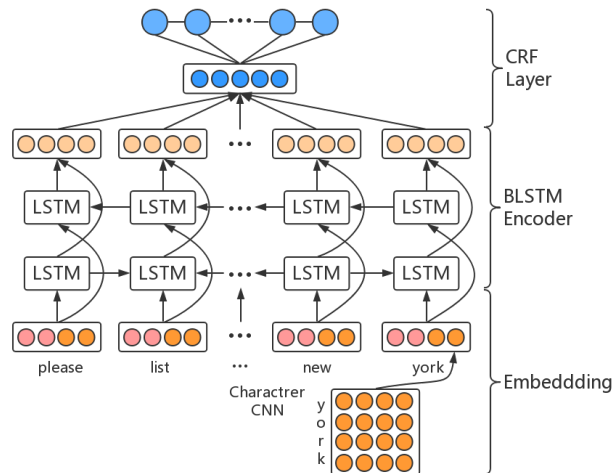


Figure 1: Architecture of CharCNN-BLSTM-CRF.

t in the label sequence; $s_j(y_t, x, t)$ is a state feature function of the label at time t and the observation sequence, i.e. scores output by the BLSTM in our case; λ and μ are parameters to be estimated, where μ is parameters of the BLSTM layer. The conditional probability distribution defined by the CRF is then:

$$p(y|x) = \frac{\exp F(y, x)}{\sum_y \exp F(y, x)} \quad (2)$$

We train the CRF by maximizing the log-likelihood of a given training set $T = \{(x_m, y_m)\}_{m=1}^M$:

$$L = \sum_m \log p(y_m|x_m) \quad (3)$$

The most probable label sequence for an input sequence is:

$$\hat{y} = \arg \max_y p(y|x) = \arg \max_y F(y, x) \quad (4)$$

Training and decoding can be solved efficiently by adopting dynamic programming algorithms.

2.4. CharNN-BLSTM-CRF

Given an utterance sequence, character-level representations are obtained by the neural network (NN), e.g. CNN, then concatenated with pre-trained word embeddings for each word in the utterance. The joint-representations are fed into the BLSTM to model the total sequence, while dropout operated on both input and output of the BLSTM is a regularization approach. Finally, the CRF layer decodes the optimal taggers. Figure 1 illustrates the architecture introduced for slot filling in detail.

3. Multi-task and cross-lingual training

3.1. Multi-task training with attention

For joint modeling of intent determination and slot filling, additional intent classification layer shares the same word-level encoder with the slot filling layer, where outputs of the BLSTM represent the entire input utterance. While not all words contribute equally to the semantic meaning of an utterance, we further introduce the attention mechanism [25] to extract such words that significantly characterize an utterance. Specifically, we first feed the outputs of the BLSTM h through a single-layer

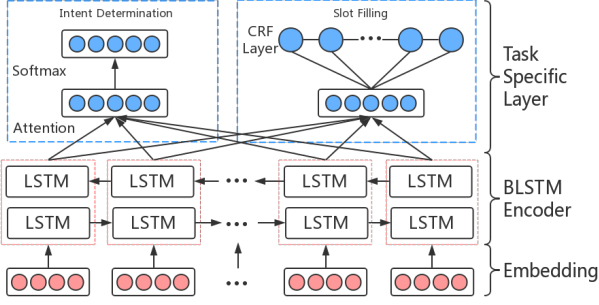


Figure 2: Joint-training neural architecture.

feed-forward network to get a hidden representation e of the entire state outputs, then obtain a normalized importance weight α_t for each state through a softmax function over e .

$$e_t = \tanh(\mathbf{W}_w h_t + b_w) \quad (5)$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_i \exp(e_i)} \quad (6)$$

$$c = \sum_t \alpha_t h_t \quad (7)$$

The context vector c of the utterance indicates part of the source sequence that should be paid attention to. We formulate it as a multi-task joint learning problem seeking to optimize a joint loss function that combines losses from different tasks. Figure 2 illustrates the proposed method for joint modeling.

3.2. Cross-lingual training

Under the assumption that reliable and clean corpora are available, this section focuses on exploring multi-level transfer learning from different languages under the previously proposed framework, which characterizes utterances hierarchically.

The first proposed approach is based on languages with similar alphabets but not restricted to be in the same language family. We share character-level representations for potential benefits that languages with similar alphabets may have some relationship in morphology and phonetics, while word-level representations are specific for each language.

Higher-level semantic representations can also be shared and transferred. A shared BLSTM encodes input-embeddings in different languages, conveys language-general information on the sequence level. Each language still maintains its own BLSTM that carries language-specific knowledge, namely specific BLSTM correspondingly. The two BLSTMs for each language are further gated and connected in the following form:

$$h_t^{combined} = f(h_t^{specific}, h_t^{shared}) \quad (8)$$

$$g_t = \sigma(\mathbf{W} h_t^{combined} + b) \quad (9)$$

$$h_t^{gated} = g_t \cdot h_t^{combined} + (1 - g_t) \cdot h_t^{specific} \quad (10)$$

where f is the combined function e.g. summation over vectors in this paper, g_t is the transform gate and $(1 - g_t)$ is the carry gate, which is inspired from highway network [26]. The gated connection adaptively carries information between the language-specific representations and language-combined representations which simultaneously characterize variant and invariant information of a language. This architecture is rather applicable for non-parallel corpora for its separability and adaptivity. The entire neural architecture for cross-lingual multi-task training is illustrated in Figure 3. Learning is performed separately via optimizing objectives for different languages.

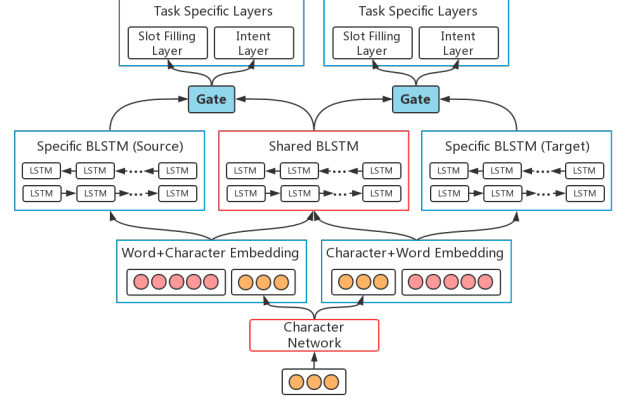


Figure 3: Cross-lingual multi-task architecture.

4. Experiment

4.1. Data

We evaluate the proposed approaches on the Airline Travel Information Systems (ATIS) dataset, a corpus in English consisting of semantically annotated spoken queries in standard BIO format. We utilize the same corpus as [12, 15, 16], where 4978 utterances are used for training and 893 utterances for testing. MIT Restaurant Corpus [27] is used as a non-parallel corpus, consisting of 7660 and 1521 utterances for training and testing respectively. However, as the latter corpus is not provided with intent annotations, we apply it to slot filling only.

In order to facilitate the research of cross-lingual SLU, we first translate the ATIS corpus from English to Spanish and German with Microsoft Translator [28], then automatically map the slot taggers from English to target languages with alignment tools. We further manually fix obvious translation errors and make the translated utterances as literal as possible. Missing and incorrect annotations are manually annotated as well.

4.2. Implementation details

In this work, we only use lexical features, regardless of hand-crafted features such as name entity types in previous models [8, 9], since they are not generally available and usually highly correlated with slot tags. GloVe embeddings trained on 6 billion words [29] are used as source language (i.e. English) word embeddings, while for target languages (i.e. Spanish and German) we take FastText [30] embeddings instead. English word embeddings are 100 dimensions while Spanish and German embeddings are 300 dimensions unless otherwise specified, all of which will be fine-tuned during training. Character embeddings in all experiments are 100 dimensions by default.

Optimization is performed with Stochastic Gradient Descent with momentum 0.9 and mini-batch size of 20. Learning decay is adopted as $\eta_t = \eta_0 / (1 + \rho t)$ (where initial rate $\eta_0 = 0.015$, decay rate $\rho = 0.05$, t is the current epoch) to prevent over-fitting, and early stopping is applied to 30 epochs. We set dropout rates to 0.5 in all experiments. Hyper-parameters for the neural architecture is similar to [12], with all states of LSTM cells set to 200, and for the CNN we use 30 filters with window length 3. Finally, we found that 100-dimension context vector for attention can achieve a better performance.

4.3. Basic architecture and joint training experiments

We first validate the effectiveness of our approaches for slot filling by the $F1$ score in turn. As shown in Table 1, additional CRF layer significantly improves the performance by 1.53%. Character-level BLSTM with 200-dimension state size is adopted to compare with the CNN. Though both CNN and BLSTM achieve better performance than word-embedding-only models, we find CNN can better characterize character-level information, achieves the best $F1$ score of 95.81%.

Table 1: Comparison of slot filling model selection.

Model	$F1$ score(%)
BLSTM	94.17
BLSTM+CRF	95.70
BLSTM+CRF+CharBLSTM	95.75
BLSTM+CRF+CharCNN	95.81

By minimizing the joint loss of intent classification and slot filling over CharCNN-BLSTM-CRF, the $F1$ score increases from 95.81% to 95.95%. Table 2 lists the error rate for intent determination and $F1$ for slot filling on several state-of-the-art models trained on the same dataset, utilizing lexical features only. With $F1$ reaches 95.95% and error rate drops down to 1.23%, we obtain state-of-the-art performance in both tasks.

Table 2: Comparison among state-of-the-art joint models.

Model	Slot(%)	Intent(%)
CNN-CRF [31]	95.42	5.91
BGRU-CRF [9]	95.49	1.90
Attention Encoder-Decoder [15]	95.87	1.57
Our Joint Model	95.95	1.23

4.4. Cross-lingual experiments

The upper half of Table 3 lists the results of different strategies for cross-lingual neural architecture on English, Spanish and German ATIS datasets. In this part, we set mono-lingual CharCNN-BLSTM-CRF with joint training as the baseline for each language, where word embeddings in all languages are set to 300 dimensions to facilitate the architecture of the shared BLSTM. Model-I utilizes shared character representations instead of language-specific character embeddings in the baseline, while 100-dimension word-embeddings are adopted for source language in Model-I* instead of that in Model-I. Model-II further incorporates shared-representations on both character-level and encoder-level as illustrated in Figure 3.

As shown in the upper half of Table 3, Model-I* that shares character representations and utilizes appropriate word-embeddings significantly outperforms the baseline in both target and source languages, with absolute improvement from 0.20% to 0.24% on slot filling, and decline of 0.11% to 0.68% on intent error rate in target languages. Specifically, jointly training with the Spanish corpus obtains the best $F1$ score of 96.17% in the English corpus. Whereas effective, this approach is sensitive to the dimension of word representations, as Model-I* generally outperforms Model-I in both target and source languages where only source word-embedding dimensions differ. Models with shared-BLSTM achieve slightly better

performance on average. When compared to Model-I, Model-II obtains comparable performance on intent and improves slot filling in the source language, but mixed performance on slot filling in target languages. Probable reason would come from the difference of word embeddings between languages.

Table 3: Cross-Lingual performance on ATIS&MIT-Restaurant.

Target Language	Model	Target		Source	
		slot	intent	slot	intent
Spanish	Baseline	93.00	1.68	95.93	1.01
	Model-I*	93.20	1.57	96.17	1.23
	Model-I	93.10	1.46	95.77	1.23
	Model-II	92.76	1.46	95.89	1.23
German	Baseline	93.61	4.93	95.93	1.23
	Model-I*	93.86	4.25	96.02	1.23
	Model-I	93.42	4.82	95.91	1.12
	Model-II	93.52	4.70	95.93	1.12
Spanish	Baseline	92.84	-	80.25	-
	Model-I	93.14	-	80.55	-
	Model-II	93.16	-	80.08	-
German	Baseline	93.38	-	80.25	-
	Model-I	93.50	-	80.21	-
	Model-II	93.54	-	80.64	-

Additional experiments on non-parallel corpus for slot filling are also illustrated in the lower half of Table 3, where ATIS for target languages and MIT Restaurant Corpus for the source language. Though mixed performance in the source language, Model-I and Model-II in target languages both achieve better performance over the baseline, the single-task SLU model, demonstrating the effectiveness and portability of our architecture to non-parallel corpora even in a different domain.

5. Conclusions

In this paper, we have demonstrated the effectiveness and scalability of the joint neural architecture for slot filling and intent determination which alleviates the OOV problem and explicitly models dependencies between neighbor labels. Assuming reliable datasets are already provided with human-supervision, which avoids MT errors and is portable to non-parallel corpora, multi-level knowledge transfer has been proposed for cross-lingual SLU. Sharing character-level representations benefited both source and target languages, while the adaptive combination of language-general and language-specific representations on the sequence-level relatively improved the performance for both languages. Portability to non-parallel corpora was further proven to be effective with the proposed architecture. Future research would focus on incorporating MT systems into this framework by taking advantage of adversarial learning, and extending to larger and more complicated data sets.

6. Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (Nos. U1536117, 11590770-4, 11504406, 11461141004) and the Key Science and Technology Project of the Xinjiang Uygur Autonomous Region (No. 2016A03007-1).

7. References

- [1] B. Jabaian, L. Besacier, and F. Lefevre, "Investigating multiple approaches for slt portability to a new language," in *INTER-SPEECH*, 2010.
- [2] F. Lefevre, F. Mairese, and S. Young, "Cross-lingual spoken language understanding from unaligned data using discriminative classification models and machine translation," in *INTER-SPEECH*, 2010.
- [3] B. Jabaian, L. Besacier, and F. Lefevre, "Combination of stochastic understanding and machine translation systems for language portability of dialogue systems," in *ICASSP*, 2011, pp. 5612–5615.
- [4] X. He, L. Deng, D. Hakkani-Tur, and G. Tur, "Multi-style adaptive training for robust cross-lingual spoken language understanding," in *ICASSP*, 2013, pp. 8342–8346.
- [5] K. Yao, G. Zweig, M.-Y. Hwang, Y. Shi, and D. Yu, "Recurrent neural networks for language understanding," in *INTERSPEECH*, 2013, pp. 2524–2528.
- [6] G. Mesnil, X. He, L. Deng, and Y. Bengio, "Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding," in *INTERSPEECH*, 2013, pp. 3771–3775.
- [7] J. D. Lafferty, A. D. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001.
- [8] G. Mesnil, Y. Dauphin, K. Yao, Y. Bengio, L. Deng, D. Hakkani-Tur, X. He, L. Heck, G. Tur, D. Yu *et al.*, "Using recurrent neural networks for slot filling in spoken language understanding," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 530–539, 2015.
- [9] X. Zhang and H. Wang, "A joint model of intent determination and slot filling for spoken language understanding," in *IJCAI*, 2016.
- [10] E. Simonnet, S. Ghannay, N. Camelin, Y. Estève, and R. De Mori, "Asr error management for improving spoken language understanding," in *INTERSPEECH*, 2017, pp. 3329–3333.
- [11] A. Celikyilmaz, R. Sarikaya, D. Hakkani-Tür, X. Liu, N. Ramesh, and G. Tür, "A new pre-training method for training deep learning models with application to spoken language understanding," in *INTERSPEECH*, 2016, pp. 3255–3259.
- [12] A. Jaech, L. Heck, and M. Ostendorf, "Domain adaptation of recurrent neural networks for natural language understanding," in *INTERSPEECH*, 2016, pp. 690–694.
- [13] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-aware neural language models," in *AAAI*, 2016, pp. 2741–2749.
- [14] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional lstm-cnns-crf," in *ACL*, vol. 1, 2016, pp. 1064–1074.
- [15] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling," in *INTER-SPEECH*, 2016, pp. 685–689.
- [16] M. Ma, K. Zhao, L. Huang, B. Xiang, and B. Zhou, "Jointly trained sequential labeling and classification by sparse attention neural networks," in *INTERSPEECH*, 2017, pp. 3334–3338.
- [17] S. Zhu and K. Yu, "Encoder-decoder with focus-mechanism for sequence labelling based spoken language understanding," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 5675–5679.
- [18] Z. Yang, R. Salakhutdinov, and W. W. Cohen, "Transfer learning for sequence tagging with hierarchical recurrent networks," *arXiv preprint arXiv:1703.06345*, 2017.
- [19] J.-K. Kim, Y.-B. Kim, R. Sarikaya, and E. Fosler-Lussier, "Cross-lingual transfer learning for pos tagging without cross-lingual resources," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2832–2838.
- [20] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning*, 2014, pp. 1764–1772.
- [21] L. Galescu, "Recognition of out-of-vocabulary words with sub-lexical language models," in *Eighth European Conference on Speech Communication and Technology*, 2003.
- [22] F. Eyben, M. Wöllmer, B. Schuller, and A. Graves, "From speech to letters-using a novel neural network architecture for grapheme based asr," in *IEEE Workshop on Automatic Speech Recognition & Understanding*, 2009, pp. 376–380.
- [23] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 2003, pp. 134–141.
- [24] H. M. Wallach, "Conditional random fields: An introduction," *Technical Reports (CIS)*, p. 22, 2004.
- [25] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1480–1489.
- [26] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015, pp. 2377–2385.
- [27] The mit restaurant corpus. [Online]. Available: <https://groups.csail.mit.edu/sls/downloads/restaurant/>
- [28] Microsoft translation online service. [Online]. Available: <https://www.bing.com/translator/>
- [29] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
- [30] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 2, 2017, pp. 427–431.
- [31] P. Xu and R. Sarikaya, "Convolutional neural network based triangular crf for joint intent detection and slot filling," in *ASRU*, 2013, pp. 78–83.