Dual Supervised Learning for Natural Language Understanding and Generation

Shang-Yu Su   Chao-Wei Huang   Yun-Nung Chen
Department of Computer Science and Information Engineering
National Taiwan University
{f05921117,r07922069}@ntu.edu.tw y.v.chen@ieee.org

Abstract

Natural language understanding (NLU) and natural language generation (NLG) are both critical research topics in the NLP field. Natural language understanding is to extract the core semantic meaning from the given utterances, while natural language generation is opposite, of which the goal is to construct corresponding sentences based on the given semantics. However, such dual relationship has not been investigated in the literature. This paper proposes a new learning framework for language understanding and generation on top of dual supervised learning, providing a way to exploit the duality. The preliminary experiments show that the proposed approach boosts the performance for both tasks.

1 Introduction

Spoken dialogue systems that can help users solve complex tasks such as booking a movie ticket have become an emerging research topic in artificial intelligence and natural language processing areas. With a well-designed dialogue system as an intelligent personal assistant, people can accomplish certain tasks more easily via natural language interactions. The recent advance of deep learning has inspired many applications of neural dialogue systems (Wen et al., 2017; Bordes et al., 2017; Dhingra et al., 2017; Li et al., 2017). A typical dialogue system pipeline can be divided into several parts: 1) a speech recognizer that transcribes a user's speech input into texts, 2) a natural language understanding module (NLU) that classifies the domain and associated intents and fills the slot values to form a semantic frame (Chi et al., 2017; Chen et al., 2017; Zhang et al., 2018; Su et al., 2018c, 2019), 3) a dialogue state tracker (DST) that predicts the current dialogue state in the multi-turn conversations, 4) a dialogue policy that determines the system action for the next step given the current state (Peng et al., 2018; Su et al., 2018a), and 5) a natural language generator (NLG) that outputs a response given the input semantic frame (Wen et al., 2015; Su et al., 2018b; Su and Chen, 2018).

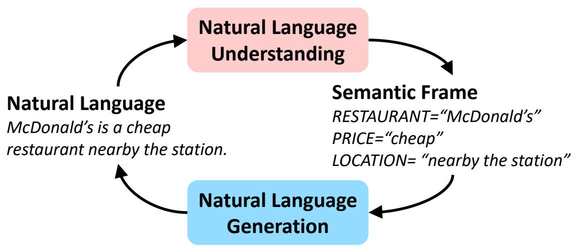

Figure 1: NLU and NLG emerge as a dual form.

Many artificial intelligence tasks come with a dual form; that is, we could directly swap the input and target of a task to formulate another task. Machine translation is a classic example (Wu et al., 2016); for example, translating from English to Chinese has a dual task of translating from Chinese to English; automatic speech recognition (ASR) and text-to-speech (TTS) also have structural duality (Tjandra et al., 2017), and so on. Previous work first exploited the duality of the task pairs and proposed supervised (Xia et al., 2017) and unsupervised (reinforcement learning) (He et al., 2016) training schemes. The recent studies magnified the importance of the duality by revealing exploiting it could boost the performance of both tasks.

NLU is to extract the core semantic concept from the given utterances, while the goal of NLG is to construct corresponding sentences based on given semantics. In other words, understanding and generating sentences are a dual problem pair shown in Figure 1. In this paper, we introduce a new training framework for NLU and NLG based on dual supervised learning (Xia et al., 2017). The experiments show that the proposed approach improves the performance for both tasks.

2 Proposed Framework

This section first describes the problem formulation, and then introduces the core training algorithm along with the proposed methods of estimating data distribution.

Given n data pairs $\{(x_i, y_i)\}_{i=1}^{n}$, the goal of NLG is to generate corresponding utterances based on given semantics. In other words, the task is to learn a mapping function $f(x; \theta_{x\to y})$ to transform semantic representations into natural language. On the other hand, NLU is to capture the core meaning of utterances, finding a function $g(y; \theta_{y\to x})$ to predict semantic representations given natural language. A typical strategy of these optimization problems is based on maximum likelihood estimation (MLE) of the parameterized conditional distribution by the learnable parameters $\theta_{x\to y}$ and $\theta_{y\to x}$.

2.1 Dual Supervised Learning

Considering the duality between two tasks in the dual problems, it is intuitive to bridge the bidirectional relationship from a probabilistic perspective. If the models of two tasks are optimal, we have *probabilistic duality*:

$$\begin{aligned} P(x)P(y \mid x; \theta_{x\to y}) &= P(y)P(x \mid y; \theta_{y\to x}) \\ &= P(x, y)\ \forall x, y, \end{aligned}$$

where $P(x)$ and $P(y)$ are marginal distributions of data. The condition reflects parallel, bidirectional relationship between two tasks in the dual problem. Although standard supervised learning with respect to a given loss function is a straightforward approach to address MLE, it does not consider the relationship between two tasks.

Xia et al. (2017) exploited the duality of the dual problems to introduce a new learning scheme, which explicitly imposed the empirical probability duality on the objective function. The training strategy is based on the standard supervised learning and incorporates the probability duality constraint, so-called *dual supervised learning*. Therefore the training objective is extended to a multi-objective optimization problem:

$$\begin{cases} \min_{\theta_{x\to y}}(\mathbb{E}[l_1(f(x; \theta_{x\to y}), y)]), \\ \min_{\theta_{y\to x}}(\mathbb{E}[l_2(g(y; \theta_{y\to x}), x)]), \\ \text{s.t. } P(x)P(y \mid x; \theta_{x\to y}) = P(y)P(x \mid y; \theta_{y\to x}), \end{cases}$$

where $l_{1,2}$ are the given loss functions. Such constraint optimization problem could be solved by introducing Lagrange multiplier to incorporate the constraint:

$$\begin{cases} \min_{\theta_{x\to y}}(\mathbb{E}[l_1(f(x; \theta_{x\to y}), y)] + \lambda_{x\to y} l_{duality}), \\ \min_{\theta_{y\to x}}(\mathbb{E}[l_1(g(y; \theta_{y\to x}), x)] + \lambda_{y\to x} l_{duality}), \end{cases}$$

where $\lambda_{x\to y}$ and $\lambda_{y\to x}$ are the Lagrange parameters and the constraint is formulated as follows:

$$\begin{aligned} l_{duality} = (&\log \hat{P}(x) + \log P(y \mid x; \theta_{x\to y}) \\ &- \log \hat{P}(y) - \log P(x \mid y; \theta_{y\to x}))^2. \end{aligned}$$

Now the entire objective could be viewed as the standard supervised learning with an additional regularization term considering the duality between tasks. Therefore the learning scheme is to learn the models by minimizing the weighted combination of original loss term and regularization term. Note that the true marginal distribution of data $P(x)$ and $P(y)$ are often intractable, here we replace them by approximate empirical marginal distribution $\hat{P}(x)$ and $\hat{P}(y)$.

2.2 Distribution Estimation as Autoregression

The current problem is how to estimate the empirical marginal distribution $\hat{P(.)}$, which arises in the reason that different data types have different structural natures. For example, natural language has sequential structures and temporal dependencies, while other structure data may not. Therefore we design an individual method of estimating distribution for each data type.

From the probabilistic perspective, we can decompose any data distribution $p(x)$ into the product of its nested conditional probability,

$$p(x) = \prod_{d}^{D} p(x_d \mid x_1, ..., x_{d-1}), \tag{1}$$

where x could be any data type and d is the index of a variable unit.

2.2.1 Language Modeling

Natural language has an intrinsic sequential nature; therefore it is intuitive to leverage the autoregressive property to learn a language model. In this work, we learn the language model based on recurrent neural networks (Mikolov et al., 2010; Sundermeyer et al., 2012) by the cross entropy objective in an unsupervised manner.

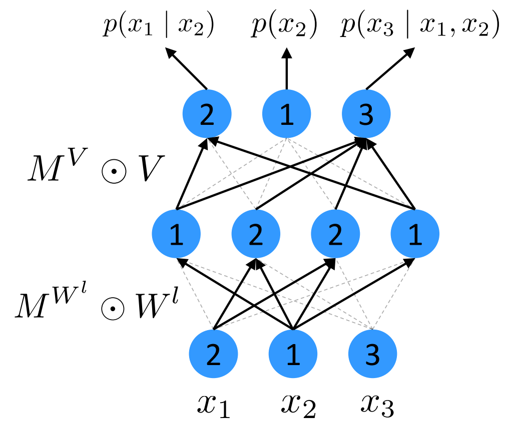

Figure 2: The masked autoencoder for distribution estimation.

2.2.2 Masked Autoencoder

The semantic representation x in our work is discrete semantic frames containing specific slots and corresponding values. Each semantic frame contains the core concept of a certain sentence, for example, the slot-value pairs "`name[Bibimbap House], food[English], priceRange[moderate], area [riverside], near[Clare Hall]`" corresponds to the target sentence "*Bibimbap House is a moderately priced restaurant who's main cuisine is English food. You will find this local gem near Clare Hall in the Riverside area.*". Even though the product rule (1) enables us to decompose any probability distribution into a product of a sequence of conditional probability, how we decompose the distribution reflects a specific physical meaning. For example, language modeling outputs the probability distribution over vocabulary space of i-th word y_i by only taking the preceding word sequence $y_{<i}$. Natural language has the intrinsic sequential structure and temporal dependency, so modeling the joint distribution of words in a sequence by such autoregressive property is logically reasonable. However, slot-value pairs in semantic frames do not have a single directional relationship between them, while they parallel describe the same sentence, so treating a semantic frame as a sequence of slot-value pairs is not suitable. Furthermore, slot-value pairs are not independent, because the pairs in a semantic frame correspond to the same individual utterance. For example, French food would probably costs more. Therefore the correlation should be taken into account when estimating the joint distribution.

Considering the above issues, to model the joint distribution of flat semantic frames, various dependencies between slot-value semantics should be leveraged. In this work, we propose to utilize a masked autoencoder (Germain et al., 2015) to estimate the marginal distribution. By zeroing certain connections, we could enforce the variable unit x_d to only depend on any specific set of variables, not necessary on $x_{<d}$; eventually we could still have the marginal distribution by product rule:

$$p(x) = \prod_d^D p(x_d \mid S_d), \tag{2}$$

where S_d is a specific set of variable units.

In practical, we elementwise-multiply each weight matrix by a binary mask matrix M to interrupt some connections, as illustrated in Figure 2. To impose the autoregressive property, we first assign each hidden unit k an integer $m(k)$ ranging from 1 to the dimension of data $D-1$ inclusively; for the input and output layers, we assign each unit a number ranging from 1 to D exclusively. Then binary mask matrices can be built as follows:

$$M = \begin{cases} 1 & \text{if } m^l(k') \geq m^{l-1}(k) \text{ or } m^L(d) > m^{L-1}(k), \\ 0 & \text{otherwise} , \end{cases}$$

where l indicates the index of the hidden layer and L indicates the output layer. With the constructed mask matrices, the masked autoencoder is shown to be able to estimate the joint distribution as autoregression. Because there is no explicit rule specifying the exact dependencies between slot-value pairs in our data, we consider various dependencies by ensemble of multiple decomposition, that is, to sample different sets S_d.

3 Experiments

To evaluate the effectiveness of the proposed model, we conduct the experiments and analyze the results.

3.1 Settings

The E2E NLG challenge dataset (Novikova et al., 2017) is utilized in our experiments, which is a crowd-sourced dataset of 50k instances in the restaurant domain. Our models are trained on the official training set and verified on the official testing set. Each instance is a pair of a semantic frame containing specific slots and corresponding values and a associated natural language utterance with

Learning Scheme		NLU	NLG			
		F1	**BLEU**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
(a)	Baseline: Iterative training	71.14	55.05	55.37	27.95	39.90
(b)	Dual supervised learning, $\lambda = 0.1$	**72.32**	**57.16**	56.37	**29.19**	**40.44**
(c)	Dual supervised learning, $\lambda = 0.01$	72.08	55.07	55.56	28.42	40.04
(d)	Dual supervised learning, $\lambda = 0.001$	71.71	56.17	55.90	28.44	40.08
(e)	Dual supervised learning w/o MADE	70.97	55.96	**59.93**	28.74	39.98

Table 1: The NLU performance reported on micro-F1 and the NLG performance reported on BLEU, ROUGE-1, ROUGE-2, and ROUGE-L of models (%).

the given semantics. The data preprocessing includes trimming punctuation marks, lemmatization, and turning all words into lowercase.

Although the original dataset is for NLG, of which the goal is to generate sentences based on the given slot-value pairs, we further formulate the NLU task as predicting slot-value pair based on the utterances, which is a multi-label classification problem. Each possible slot-value pair is treated as an individual label, and the total number of labels is 79. To evaluate the quality of the generated sequences regarding both precision and recall, for NLG, the evaluation metrics include BLEU and ROUGE (1, 2, L) scores with multiple references, while F1 score is measured for NLU results.

3.2 Model

The model architectures for NLG and NLU are gated recurrent unit (GRU) (Cho et al., 2014) with two identical fully-connected layers at two ends of the GRU. Thus the model is symmetrical and may have semantic frame representation as initial and final hidden states and sentences as the sequential input.

In all experiments, we use mini-batch *Adam* as the optimizer with each batch of 64 examples, 10 training epochs were performed without early stop, the hidden size of network layers is 200, and word embedding is of size 50 and trained in the end-to-end fashion.

3.3 Results and Analysis

The experiment results are shown in Table 1, where each reported number is averaged over the results on the official testing set from three different models. The row (a) is the baseline that trains NLU and NLG separately and independently, and the rows (b)-(d) are the results from the proposed approach with different Lagrange parameters.

The proposed approach incorporates probability duality into the objective as the regularization term. To examine its effectiveness, we control the intensity of regularization by adjusting the Lagrange parameters. The results (rows (b)-(d)) show that the proposed method outperforms the baseline on all automatic evaluation metrics. Furthermore, the performance improves more with stronger regularization (row (b)).

In this paper, we design the methods for estimating marginal distribution for data in NLG and NLU tasks: language modeling is utilized for sequential data (natural language utterances), while the masked autoencoder is conducted for flat representation (semantic frames). The proposed method for estimating the distribution of semantic frames considers complex implicit dependencies between semantics by ensemble of multiple decomposition of joint distribution. In our experiments, the empirical marginal distribution is the average over the results from 10 different masks and orders; in other words, 10 types of dependencies are modeled. The row (e) can be viewed as the ablation test, where the marginal distribution of semantic frames is estimated by considering slot-value pairs independent to others and statistically computed from the training set. The performance is worse than the ones that model the dependencies, demonstrating the importance of considering the nature of input data.

4 Conclusion

This paper proposes a new training framework for natural language understanding and generation based on dual supervised learning, which exploits the duality between NLU and NLG and introduces it into the learning objective as the regularization term. Moreover, domain knowledge is incorporated to design suitable approaches to estimating data distribution. The proposed methods demonstrate effectiveness by boosting the performance of both tasks simultaneously in the benchmark experiments.

References

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *Proceedings of ICLR*.

Po-Chun Chen, Ta-Chung Chi, Shang-Yu Su, and Yun-Nung Chen. 2017. Dynamic time-aware attention to speaker roles and contexts for spoken language understanding. In *Proceedings of ASRU*.

Ta-Chung Chi, Po-Chun Chen, Shang-Yu Su, and Yun-Nung Chen. 2017. Speaker role contextual modeling for language understanding and dialogue policy learning. In *Proceedings of IJCNLP*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of EMNLP*, pages 1724–1734.

Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2017. Towards end-to-end reinforcement learning of dialogue agents for information access. In *Proceedings of ACL*, pages 484–495.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. 2015. Made: Masked autoencoder for distribution estimation. In *International Conference on Machine Learning*, pages 881–889.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828.

Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017. End-to-end task-completion neural dialogue systems. In *Proceedings of The 8th International Joint Conference on Natural Language Processing*.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.

Jekaterina Novikova, Ondrej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of SIGDIAL*, pages 201–206.

Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, Kam-Fai Wong, and Shang-Yu Su. 2018. Deep dyna-q: Integrating planning for task-completion dialogue policy learning. *arXiv preprint arXiv:1801.06176*.

Shang-Yu Su and Yun-Nung Chen. 2018. Investigating linguistic pattern ordering in hierarchical natural language generation. In *Proceedings of 7th IEEE Workshop on Spoken Language Technology*.

Shang-Yu Su, Xiujun Li, Jianfeng Gao, Jingjing Liu, and Yun-Nung Chen. 2018a. Discriminative deep dyna-q: Robust planning for dialogue policy learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Shang-Yu Su, Kai-Ling Lo, Yi-Ting Yeh, and Yun-Nung Chen. 2018b. Natural language generation by hierarchical decoding with linguistic patterns. In *Proceedings of The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Shang-Yu Su, Pei-Chieh Yuan, and Yun-Nung Chen. 2018c. How time matters: Learning time-decay attention for contextual spoken language understanding in dialogues. In *Proceedings of NAACL-HLT*.

Shang-Yu Su, Pei-Chieh Yuan, and Yun-Nung Chen. 2019. Dynamically context-sensitive time-decay attention for dialogue modeling. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7200–7204. IEEE.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2017. Listening while speaking: Speech chain by deep learning. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 301–308. IEEE.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of EACL*, pages 438–449.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yingce Xia, Tao Qin, Wei Chen, Jiang Bian, Nenghai Yu, and Tie-Yan Liu. 2017. Dual supervised learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3789–3798. JMLR. org.

Rui Zhang, Honglak Lee, Lazaros Polymenakos, and Dragomir Radev. 2018. Addressee and response selection in multi-party conversations with speaker interaction rnns. In *Proceedings of AAAI*.