When Generative Adversarial Networks Meet Sequence Labeling Challenges

Yu Tong[†] [*], **Ge Chen**[‡], **Guokai Zheng**[§], **Rui Li**[†], **Dazhi Jiang**[†]
[†] Department of Computer Science, Shantou University
[‡] Midea Group
[§] XPeng Motors
[†] {tongyu, ruili, dzjiang}@stu.edu.cn
[‡] {chenge6}@midea.com
[§] {zhenggk}@xiaopeng.com

Abstract

The current framework for sequence labeling encompasses a feature extractor and a sequence tagger. This study introduces a unified framework named SLGAN, which harnesses the capabilities of **G**enerative **A**dversarial **N**etworks to address the challenges associated with Sequence Labeling tasks. SLGAN not only mitigates the limitation of GANs in backpropagating loss to discrete data but also exhibits strong adaptability to various sequence labeling tasks. Unlike traditional GANs, the discriminator within SLGAN does not discriminate whether data originates from the discriminator or the generator; instead, it focuses on predicting the correctness of each tag within the tag sequence. We conducted evaluations on six different tasks spanning four languages, including Chinese, Japanese, and Korean Word Segmentation, Chinese and English Named Entity Recognition, and Chinese Part-of-Speech Tagging. Our experimental results illustrate that SLGAN represents a versatile and highly effective solution, consistently achieving state-of-the-art or competitive performance results, irrespective of the specific task or language under consideration. [1]

1 Introduction

Generative Adversarial Networks (GANs) have achieved remarkable success in the realm of computer vision. Nonetheless, their applicability in the field of Natural Language Processing (NLP) has been limited due to the discrete nature of natural language text. This discretization challenge arises from the inability to directly back-propagate gradients from the discriminator to the discrete data, thereby impeding the integration of GANs into NLP tasks. While alternative approaches, such as Reinforcement Learning (RL) (Guo, 2015; Yu et al., 2017) methodologies or smooth approximation policies (Zhang et al., 2016), have been proposed to circumvent this limitation, the predominant utilization of GANs in NLP has been confined to text generation tasks, including summarization, poetry generation, dialog systems, and machine translation. Notably, there are scarce applications in sequence labeling tasks, primarily due to the inherent distribution disparities between natural language text and sequence labels. Furthermore, existing approaches for sequence labeling predominantly employ a pre-trained model and sequence tagger to model data distributions and subsequently produce sequences of tags. The efficacy of these methods is significantly constrained by the quality of the training data and the inherent limitations of the network's learning capabilities. It is well-recognized that the fundamental principle of Generative Adversarial Networks (GANs) revolves around harnessing adversarial dynamics between the discriminator and the generator, with the ultimate aim of enhancing the generator's data generation performance through the critical feedback loop provided by the discriminator. Leveraging this intrinsic strength of GANs, we introduce a GAN-based framework and introduce a specialized discriminator in a bystander role. The discriminator's role is to guide the existing sequence labeling model, with the overarching objective of elevating the quality of the training process.

We present a unified framework for sequence labeling, denoted as SLGAN, which integrates Generative Adversarial Network (GAN) principles. SLGAN enhances the training process of the sequence labeling model by incorporating a purposefully designed discriminator. The schematic representation of the SLGAN framework is shown in Figure 1, which comprises two pivotal constituents: the generator and the discriminator. We have constructed a generator, which relies on a pre-trained BERT (Kenton and Toutanova, 2019) model and a

[*] Corresponding author.

[1] Our source code has been made publicly available at https://github.com/aashenge/slgan.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 10625–10635
November 12-16, 2024 ©2024 Association for Computational Linguistics

conventional sequence tagger CRF (Lafferty et al., 2001), to produce the sequence of tags. Consequently, the loss function for the generator comprises two main components. One component stems from the intrinsic supervised information of the generator, while the other component is derived from the loss feedback originating from the discriminator. In contrast to the discriminator in the standard GAN, which is primarily responsible for distinguishing between generated and real data distributions, the discriminator in our SLGAN framework is explicitly tailored to evaluate the correctness of each tag within the tag sequence. This evaluation is carried out by observing both the textual content and the accompanying labels simultaneously. The discriminator in our framework simultaneously processes two distinct data streams: **(i)** A composite data stream consisting of textual information combined with labels generated by the generator. **(ii)** Another composite data stream comprises textual information and ground truth labels. The fusion of text and labels is facilitated through an attention layer. The loss incurred from the first data stream is back-propagated into the system to continuously refine the generator's output, striving for high-quality label generation while concurrently intensifying the complexity of the discrimination task. In summary, SLGAN leverages the supervised guidance provided by the auxiliary discriminator to train an improved generator for sequence labeling tasks.

- SLGAN realizes the integration and tailoring of the generative-adversarial mechanism to suit the sequence labeling tasks.
- SLGAN is a unified and comprehensive framework, consistently yielding enhancements across the entire spectrum of sequence labeling tasks.
- SLGAN is effective for sequence labeling tasks regardless of language and domain.

2 Related Work

Generative adversarial networks (GANs) have achieved remarkable success in the field of computer vision (Radford and Metz, 2021; Chen et al., 2016; Salimans et al., 2016) and image generation (Karras et al., 2019; Wang et al., 2018). Through the process of adversarial training, where the generator engages in a competitive learning paradigm with the discriminator, the generator generates data that closely approximates the distribution of real-world data. Nevertheless, to the best of our knowledge, the effective application of GANs in the domain of Natural Language Processing (NLP) has been limited due to the inherently discrete nature of textual data. To tackle this challenge, the work proposed by Zhang et al. employed Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks in their adversarial training approach, aimed at generating realistic textual content. Notably, they replaced the traditional GAN objective with a feature distribution matching criterion during the generator's training process. On one hand, researchers introduced a method to handle the non-differentiability of the multinomial distribution by replacing the softmax function with the Gumbel-softmax distribution, which offers a continuous approximation of the multinomial distribution (Kusner and Hernández-Lobato, 2016). On the other hand, researchers suggested furnishing the discriminator with the intermediate vector generated by the generator, as opposed to the entire sequence output (Lamb et al., 2016). It is noteworthy that GANs primarily center their applications around text generation tasks, while their participation in the domain of sequence labeling tasks remains limited.

Typically, within the sequence labeling framework, an initial feature extractor is employed, such as LSTM or CNN (Ma and Hovy, 2016) architectures. Subsequently, a sequence tagger, notably Conditional Random Field (CRF) networks (Lafferty et al., 2001), is utilized to assign tags to each character. With the advancement of pre-trained language models like BERT (Kenton and Toutanova, 2019) and ELMo (Peters et al., 2018), a gradual shift has occurred, replacing LSTM and CNN with these potent models. ELECTRA (Clark et al., 2020) was proposed in an economically efficient way, by replacing the Masked Language Modeling (MLM) task in BERT with a task focused on detecting replaced tokens. ELECTRA engaged in training a discriminator to predict whether each token is being substituted with a sample generated by the model. SLGAN draws inspiration from the ELECTRA. The distinguishing factor between SLGAN and ELECTRA lies in their objectives. SLGAN's core emphasis is on optimizing performance in sequence labeling tasks, whereas ELECTRA aims to establish a robust pre-trained model. After the pre-training phase, the authors discarded the generator, only retaining the discriminator component, which represents the ELECTRA model. Consequently, ELECTRA is a discriminator. Differing from ELECTRA, SLGAN, in line with the established paradigm of GANs and their derivative models, pursues the goal of enhancing generator quality.

3 Proposed Method

The architecture of the SLGAN framework is illustrated in Figure 1. It is important to note that two distinct BERT models are employed within the generator and discriminator components. In both instances, the role of the BERT model remains consistent, primarily focused on obtaining representations of the input sentences. The input sentences undergo independent processing within the respective BERT models, yielding two distinct streams of input representations. The initial input stream is dedicated to the task of generating tag sequences within the generator, while the second stream is specifically utilized for participating in attention operations. Given that the input sequences and the predicted labels reside within disparate data distributions, an attention layer is deployed to integrate these distinct data distributions. Specifically, the logits generated by the generator and those derived from the ground truth labels are individually merged with the original input representations via the attention layer. The logits from the generator and those from the conversion of ground truth labels are individually integrated with the original input representations through the use of the attention layer. It is noteworthy that the discriminator remains agnostic to the data's source, be it the generator or the original input data. The primary mission of the discriminator is to evaluate the accuracy of each tag within the tag sequences within both data streams.

3.1 Generative Model

The generator, located in the lower-left section of the figure, is a compact sequence labeling model composed of an encoder and a decoder. For the encoder, we employ the pre-trained BERT model, while the decoder is implemented using a CRF layer. The generator leverages the pre-trained BERT model to project the input sequences into a vector space. This process yields a feature map, which is subsequently processed through two fully connected layers, each equipped with the Rectified Linear Unit (ReLU) activation function. Following the fully connected layer, a CRF network, a widely employed sequence tagger, is utilized to generate the tag sequences. The output of the fully connected layer, denoted as $logits_1$, is simultaneously input to both the CRF and attention layers.

In Figure 1, in line with the established model for sequence labeling tasks, we designate the input as the observational sequence: $X = \{x_1, x_2, ...x_n\}$, $n \in N^+$, take the NER task as an example and employ the typical "BMESO" tagging schema, the corresponding tags are the hidden state sequence: $Y = \{y_1, y_2, ...y_n\}$ and $y_i \in \{B\_type, M\_type, E\_type, S\_type, O\}$ $\forall i \leq n, i \in N^+$. The prefix serves to denote the relative position of the present character within an entity, while the suffix "type" signifies the entity type. The overall count of tags equals $n * m + 1$, with n representing the entity types and m denoting the tag categories. $Y \in L^n$, here, L^n denotes the set of all potential tag sequences. The decoder CRF endeavors to discover the optimal tag sequence Y' based on the input X, in the following manner:

$$Y^{'} = \arg\max_{Y \in L^n} P(Y|X) \tag{1}$$

3.2 Discriminative Model

The structure of the discriminator includes a feature extractor layer, a Gumbel Softmax (Jang et al., 2017) layer, an attention layer, a Fully Connected (FC) layer, and a softmax layer. The discriminator concurrently processes three inputs: $logits_1$ and $logits_2$, which respectively correspond to the output of the generator and the transformed values derived from the ground truth Y, as well as the representation of the input sequences X.The input text is encoded through a separate BERT model to produce feature representations. Each token exhibits a feature dimension of 768, resulting in a collective feature size for the entire sentence, represented as X, with dimensions of $(768 * n)$, where n signifies the sequence length. An attention layer is applied to merge the representations of X and Y, yielding two separate attention outputs denoted as z_1 and z_2. Here, X is defined as a set of embeddings, $X = \{emb\_1, emb\_2, ..., emb\_n\}$, while Y takes values in $\{logits\_1, logits\_2\}$, and the shape of Y is $label\_size * n$. To facilitate the subsequent attention operation, X and Y are initially resized to achieve a $middle\_size$ shape. Following the attention operation, the shape of z remains consistent, retaining dimensions of $(middle\_size * n)$.

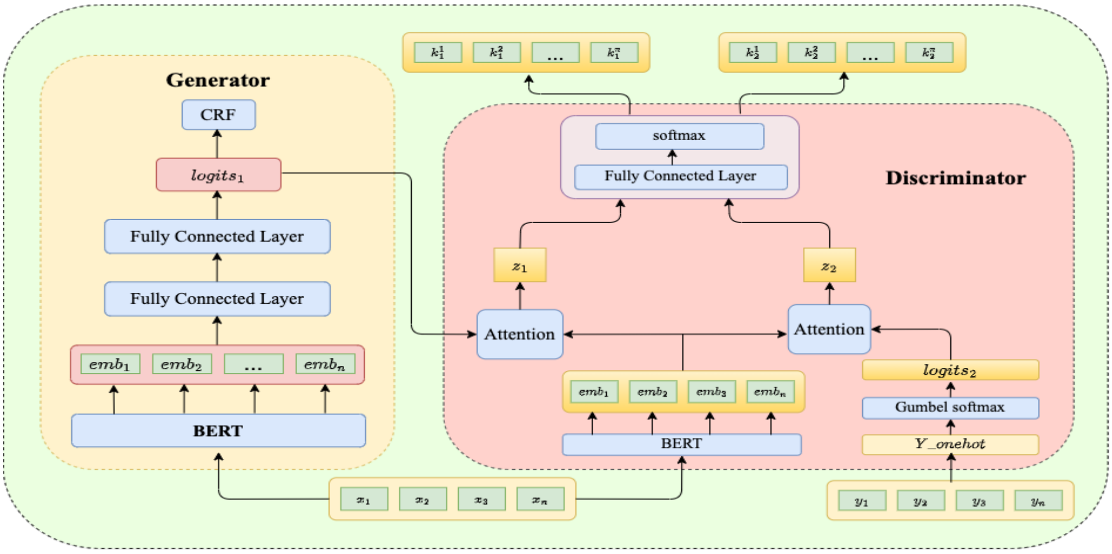


Figure 1: Overview of the SLGAN training process. The framework is rooted in the GAN structure, comprising a generator (G) and a discriminator (D). The generator (G) is a versatile model capable of generating tag sequences, with a traditional approach being adopted in our study. The pre-trained BERT model is leveraged as a feature extractor, while a CRF network is employed as a sequence tagger within the generator. On the other hand, the discriminator (D) serves as a classifier tasked with distinguishing the accuracy of each tag within the tag sequence. "D" is trained on both the original input data and the output of "G". "G" is trained using a composite objective that encompasses the supervised loss from the sequence labeling model and the loss back-propagated by "D".

Subsequently, both z_1 and z_2 are independently processed through subsequent Fully Connected (FC) and softmax layers. The FC layer serves to transform the features to achieve the appropriate size for the final classification task. The compacted features are then directed into the softmax layer. The shrunken features have dimensions of $(2 * n)$ since the primary objective of the discriminator is to assess the correctness of each tag within the tag sequence, denoted as $K = \{k^1, k^2, ..., k^n\}(k \in \text{“}right\text{”}, \text{“}wrong\text{”})$. Considering that the discriminator makes predictions based on both z_1 and z_2, the size of the output K is twice the batch size. The softmax layer employs the cross-entropy loss function for computation.

3.2.1 Gumbel Softmax Layer

Initially, the true labels are transformed from a discrete distribution into one-hot encoding, represented as Y_onehot. Given that Y_onehot lacks smoothness, we proceed to convert it into an approximately continuous and smooth distribution, denoted as $logits_2$, using the Gumbel softmax approach as in Equation 2. Subsequently, $logits_2$ is actively engaged in the attention operation with X.

$$logits\_2 = softmax(1/\tau(h + g * \alpha)) \qquad (2)$$

Here, h represents the one-hot distribution Y_onehot, g corresponds to the Gumbel-Softmax distribution, and the temperature τ is fixed at a value of 0.5. Additionally, the parameter α is specifically set to 0.1.

3.2.2 Attention Layer

The discriminator encounters a challenge when attempting to determine the correctness of each tag solely based on observing the tag sequences. Consequently, the text features serve as a necessary reference and are combined with the label features. Given that the input text and label categories inherently belong to different data distributions, an attention layer is employed to establish their association. To facilitate this integration, we employ the Scaled Dot-Product Attention mechanism to correlate the representations of $(X, logits\_1)$ and $(X, logits\_2)$, enabling the fusion of relevant information.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \qquad (3)$$

Here, we define the matrices involved in the attention operation. Specifically, Q is equivalent to X, and K is set as Y. The matrix X, representing the features generated by BERT, exhibits dimensions of $(768 * n)$, while matrix Y is structured with dimensions of $(label\_size * n)$. To facilitate their participation in the attention operation, both X and Y are resized to align with a $middle\_size$, which we have designated as having a value of 200. The dimension d_k is set to be equal to $middle\_size$, and n signifies the sequence length.

3.3 Loss Function

The generator of SLGAN follows the conventional approach used in sequence labeling tasks, wherein the CRF loss function is adopted. Simultaneously, the primary objective of the SLGAN discriminator is the accurate prediction of tags within the tag sequences. Accordingly, the widely-used cross-entropy loss function, common in sequence labeling tasks, is applied. The discriminator endeavors to minimize the cross-entropy loss. In summation, SLGAN is designed to minimize the overall loss function as described below:

$$\min_G \min_D \mathcal{L}(G, D) \tag{4}$$

$$\mathcal{L}_G = \mathcal{L}_{CRF} + \lambda \mathcal{L}_{D\_out} \tag{5}$$

$$\mathcal{L}_D = \mathcal{L}_{D\_out} + \mathcal{L}_{D\_truth} \tag{6}$$

In Eq. (5), the parameter λ plays a crucial role in weighting the contributions of the two distinct types of loss and is explicitly configured to a value of 1. The loss functions $\mathcal{L}_{D\_out}$ and $\mathcal{L}_{D\_truth}$ correspond to cross-entropy losses, which are computed from data streams originating from the generator's output and the realistic data distribution, respectively. Concurrently, $\mathcal{L}_{CRF}$ represents the CRF loss. It is essential to note that $\mathcal{L}_{D\_out}$ is subsequently backpropagated to the generator for further training.

$$\mathcal{L}_{CRF} = -\log p(y|X) = -s(X, y) + \log(\sum_{y \in Y_x} e^{s(X,y)}) \tag{7}$$

The function s denotes the scoring function, and Y_x represents all tag sequences.

4 Experiments

The evaluation encompasses a diverse set of tasks, spanning in-domain Chinese Word Segmentation (CWS), cross-domain CWS, Korean Word Segmentation (KWS), Japanese Word Segmentation (JWS), Chinese Named Entity Recognition (NER), English NER, and Chinese Part-of-Speech (PoS) Tagging. These tasks encompass four distinct languages and are evaluated across fourteen datasets. To illustrate the specifics of our experimental setup, for the CWS and Chinese NER tasks, we employ BERT-base-Chinese as the feature extractor. For JWS and KWS tasks, BERT_Multilingual is employed, serving as the feature extraction layer. For the English NER experiments, the BERT-Base-Cased is utilized. It is noteworthy that a consistent "BMES" tagging schema is applied uniformly across all experiments to ensure a standardized evaluation framework.

4.1 Datasets

4.1.1 In-domain and Cross-domain Chinese Word Segmentation

In the assessment of the CWS task, we consider five distinct datasets. Specifically, for simplified Chinese, we leverage the PKU, MSR, and Chinese Penn Treebank 6.0 (CTB6) datasets. In the case of traditional Chinese, the AS and CITYU datasets are utilized for evaluation. Our evaluation encompasses the cross-domain CWS task, encompassing five distinct datasets. These datasets are further categorized into two domains: the Chinese fantasy novel domain and the patent domain. Within the realm of Chinese fantasy novels, we consider the DoLuoDaLu (DL), FanRenXiuXianZhuan (FR), and ZhuXian (ZX) datasets (Qiu and Zhang, 2015). Simultaneously, datasets originating from the dermatology and patent domains, namely Dermatology (DM) and Patent (PT) (Ye et al., 2019), are included. It is essential to highlight that these datasets are drawn from domains characterized by the emergence of numerous neologisms, especially originating from fiction, thereby exacerbating the out-of-vocabulary (OOV) issue. In conducting our evaluation, we adhere to the dataset configurations detailed in the work (Ye et al., 2019) to ensure a comprehensive and consistent assessment.

4.1.2 Japanese and Korean Word Segmentation

In evaluating the JWS task, we rely on Version 1.1 of the extensively employed Balanced Corpus of Contemporary Written Japanese (BCCWJ) dataset (Maekawa et al., 2014). To ensure a systematic evaluation, we perform a dataset split in accordance with the guidelines provided by the Project Next NLP. Our evaluation of the KWS task encompasses two distinct datasets: the UD_Korean-GSD corpora [2] and the Kaist [3] datasets. These datasets were originally purposed for assessing syntactic parsing tasks and are derived from the KAIST

[2] https://github.com/emorynlp/ud−korean/tree/master/-google

[3] https://github.com/UniversalDependencies/UD_Korean-Kaist

Treebank (Choi et al., 1994) and the Google UD Treebank (McDonald et al., 2013). To construct these datasets, we extract words guided by syntactic information that implicitly conveys word boundaries. Furthermore, the dataset split is performed in accordance with the Universal Dependencies in the Korean (UDK) project to maintain consistency and alignment with established standards.

4.1.3 Chinese and English Name Entity Recognition

In our evaluation of the Chinese NER task, we consider four datasets: OntoNotes4.0 (Pradhan et al., 2011), Weibo (Peng and Dredze, 2015), MSRA (Levow, 2006), and Resume (Zhang and Yang, 2018). For the English NER task, the CoNLL2003 dataset (Tjong Kim Sang and De Meulder, 2003) is employed as the benchmark. This comprehensive selection of datasets ensures a rigorous evaluation across multiple languages and domains.

4.1.4 Chinese Part-of-Speech Tagging

In the assessment of the Chinese PoS Tagging task, we employ four distinct Chinese datasets for the evaluation of SLGAN. These datasets encompass CTB5, CTB6, and CTB9, originating from the Penn Chinese TreeBank (Xue et al., 2005), and UD1 (Universal Dependencies) (Nivre et al., 2016). CTB5, CTB6, and CTB9 are in simplified Chinese, while UD1 is in traditional Chinese. To maintain methodological consistency, we adhere to the official dataset splits as prescribed in our experiments. This comprehensive dataset selection facilitates a comprehensive evaluation of SLGAN's performance across different languages and domains.

4.2 Parameter Settings and Evaluation Metrics

In our experimental settings, we establish a consistent sequence length of 128 across all tasks. The batch size is set at 64, and we conduct training for 15, 10, and 10 epochs for the NER, WS, and PoS Tagging tasks, respectively. To optimize the training process, we employ the Adam optimizer with a learning rate of $2e-5$. To mitigate overfitting, we incorporate an early stop mechanism into our training process. Beyond the conventional Micro F1 score, we recognize the significance of the Out-of-Vocabulary (OOV) error as a crucial metric for assessing a word segmentation model's generalization capacity. Consequently, we introduce OOV recall (R_oov) to evaluate the generalization ability of SLGAN in the WS task. Furthermore, to ascertain the model's reliability and consistency, we monitor the Standard Deviation (SD) values across multiple experiments. This practice helps ensure the robustness and stability of the SLGAN model.

4.3 Experimental Results

The evaluation results for the WS, NER, and PoS Tagging tasks are presented in Tables from Table 1 to Table 6. To enhance readability, the results have been rounded to one or two decimal places for precision and clarity.

4.3.1 Results of In-Domain Chinese Word Segmentation

In Table 1, we present the results of the in-domain CWS task. SLGAN achieves state-of-the-art (SOTA) F1 scores across all five datasets. Notably, in addition to the F1 score, SLGAN significantly enhances the recall of Out-of-Vocabulary (OOV) words. SLGAN attains a new SOTA performance in terms of R_oov scores on all five datasets, which encompass both simplified and traditional Chinese. Impressively, SLGAN showcases an outstanding improvement of **+10.47%** in R_oov score on the AS dataset compared to existing dominant methods. These notable enhancements in OOV recall underscore the robustness and exceptional generalization ability of SLGAN.

4.3.2 Results of Cross-Domain Chinese Word Segmentation

Table 2 provides a comprehensive overview of the evaluation results in the cross-domain CWS task. The first two rows represent results reported by Liu et al. and Zhang et al., while it's important to note that these methods were not evaluated on all five datasets. Therefore, we refer to the evaluation results reported by Ding et al. for comparison. SLGAN demonstrates its robust performance on cross-domain Chinese CWS without the need for additional data from the target domain, setting it apart from the Partial-CRF model (Liu et al., 2014), which relies on partially labeled data and combines data from various sources into a unified format. In comparison to methods (Zhang et al., 2018; Ye et al., 2019), SLGAN achieves competitive results without the use of extra dictionaries or word embeddings specifically trained for the target domain. Notably, when compared with the four methods exclusively developed for cross-domain CWS, SLGAN

	CITYU		AS		PKU		MSR		CTB6	
	F1	R_oov	F1	R_oov	F1	R_oov	F1	R_oov	F1	R_oov
Gong et al., 2019	96.2	73.58	95.2	77.33	96.2	69.88	97.8	64.2	97.3	83.89
Huang et al., 2020	97.6	87.27	96.6	79.26	96.6	79.71	97.9	83.35	**97.6**	87.77
Meng et al., 2019	**97.9**	-	**96.7**	-	**96.7**	-	**98.3**	-	-	-
Tian et al., 2020	97.8	**87.57**	96.58	78.48	96.51	**86.76**	98.28	**86.67**	97.16	88.00
BERT_CRF	97.61	85.98	96.50	**83.36**	96.50	79.78	98.10	86.54	96.58	**88.98**
SLGAN	**97.82**	**87.71**	**96.63**	**89.73**	**96.70**	**94.22**	**98.30**	**96.28**	**97.70**	**90.72**
SLGAN (SD)	0.24	0.12	0.13	0.26	0.39	0.51	0.26	0.18	0.25	0.12

Table 1: Comparison of SLGAN with established primary methods in the CWS task. We assess the performance using the F1 score and R_oov as evaluation metrics. SLGAN (SD) represents the Standard Deviation obtained from five experiments.

	DL	FR	ZX	DM	PT
Liu et al., 2014	92.5	90.2	83.9	82.8	85.0
Zhang et al., 2018	92.0	89.1	88.8	81.2	85.9
Ye et al., 2019	93.5	89.6	89.6	82.2	85.1
Ding et al., 2020	**94.1**	**93.1**	90.9	85.0	89.6
Tong et al., 2022a	92.1	90.8	**90.9**	88.4	**92.37**
SLGAN	93.29	91.76	90.27	**87.79**	**90.83**
SLGAN (SD)	0.24	0.2	0.36	0.25	0.13

Table 2: Comparison between SLGAN and established primary methods in the Cross-Domain CWS task. We evaluate performance using the F1 score as the primary metric.

	BCCWJ	
	F1	R_oov
Kitagawa and Komachi, 2018	98.42	-
Higashiyama et al., 2019	98.93	-
Tong et al., 2022b	**98.94**	**93.01**
BERT_CRF	97.71	90.08
SLGAN	98.5	**95.19**
SLGAN (SD)	0.08	0.12

Table 3: Comparison between SLGAN and established primary methods in the JWS task. The metrics are the F1 score and R_oov.

requires no additional optimization efforts. For fair and comparable evaluations, SLGAN solely trains a base segmenter using the PKU dataset. On the DM dataset, SLGAN even surpasses the SOTA results, achieving a notable **+2.79%** increase in F1 score. Additionally, there is a **+1.23%** boost on the PT dataset. These substantial improvements in the cross-domain CWS task serve as strong evidence of SLGAN's impressive generalization capabilities.

4.3.3 Results of Japanese Word Segmentation

Table 3 presents the evaluation outcomes for Japanese Word Segmentation (JWS). SLGAN achieves results close to the SOTA without relying on word dictionaries or character type information, as employed in prior works (Higashiyama et al., 2019; Kitagawa and Komachi, 2018). In comparison to the BERT_CRF model, SLGAN delivers a notable increase of **+1.04%** in the F1 score, highlighting its enhanced performance. SLGAN also demonstrates significant proficiency in handling out-of-vocabulary (OOV) words in the Japanese dataset. These evaluation results provide further evidence of SLGAN's effectiveness, particularly in dealing with OOV words, and underscore its robust performance in the Japanese context.

	GSD		KAIST	
	F1	R_oov	F1	R_oov
Tong et al., 2022b	92.37	83.81	91.19	82.24
BERT_CRF	87.12	78.27	87.62	78.34
SLGAN	**87.57**	**85.14**	**88.90**	**88.70**
SLGAN (SD)	0.22	0.18	0.24	0.12

Table 4: Comparison of SLGAN with established methods in the KWS task. F1 score and R_oov are the metrics.

4.3.4 Results of Korean Word Segmentation

For the KWS task, no directly related baselines are available, as the Korean datasets we utilized are typically designed for assessing syntactic parsing. To provide a meaningful comparison, we contrast SLGAN's performance with that of the BERT_CRF model and perform additional experiments. Table 4 outlines the evaluation results of SLGAN on the KWS task. Notably, the KAIST dataset sees a substantial increase of **+1.31%** in F1 score. Furthermore, there is a remarkable **+10.88%** enhancement in OOV word recall. These results underscore SLGAN's effectiveness in addressing the KWS task, particularly in enhancing OOV word recall.

4.3.5 Results of Chinese and English NER

Table 5 displays the results of the Chinese and English NER tasks. The evaluation results are rounded to two decimal places. SLGAN achieves near SOTA results. While the FGN model (Xuan et al., 2020) has achieved SOTA performance on the Weibo and Resume datasets, FGN leverages glyph information, which is particularly useful for

	OnNote	Weibo	MSRA	Resume	CoNLL
Li et al., 2020	**84.47**	-	**96.72**	-	93.33
Wu et al., 2021	82.57	70.43	96.24	95.98	-
Zhu and Li, 2022	82.83	**72.66**	96.26	**96.66**	**93.65**
Xiong et al., 2023	81.47	68.23	95.42	-	-
Yang et al., 2023	82.66	71.94	-	96.2	-
BERT_CRF	79.16	67.33	94.80	95.78	91.08
SLGAN	82.30	71.05	96.10	96.63	93.29
SLGAN (SD)	0.16	0.20	0.11	0.25	0.18

Table 5: SLGAN v.s. existing primary methods of the NER task. The F1 score is the metric.

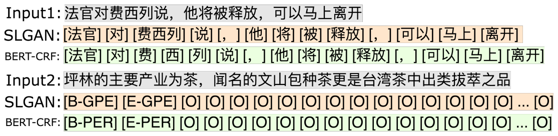
	CTB5	CTB6	CTB9	UD1
(Meng et al., 2019)	96.61	95.41	**93.15**	96.14
(Liu et al., 2021)	97.14	95.18	-	96.06
(Li et al., 2020)	**97.92**	**96.57**	-	**96.98**
BERT_CRF	96.06	94.77	92.29	94.79
SLGAN	96.78	94.86	**94.61**	96.11
SLGAN (SD)	0.19	0.14	0.18	0.06

Table 6: Comparison of SLGAN with established methods in the Chinese PoS Tagging task. We employ the F1 score as the metric.

Chinese characters due to their pictographic nature. In contrast, SLGAN relies solely on text features. The BERT-MRC+DSC (Li et al., 2020) surpasses our SLGAN on the OntoNotes 4.0 dataset. However, it's essential to note that BERT-MRC+DSC is based on the Machine Reading Comprehension (MRC) framework, which involves introducing substantial external knowledge, such as synonyms, and formalizing the NER task as an MRC task. Additionally, BERT-MRC+DSC utilizes dice loss as a replacement for conventional cross-entropy loss to address data imbalance issues. SLGAN achieves competitive performance without relying on external information. In both tasks, SLGAN is trained solely on its training splits. Even without specific optimizations, SLGAN achieved near SOTA results in Chinese NER. It secured the SOTA result on the MSRA dataset and also performed remarkably well in English NER. Comparing SLGAN to BERT alone, we observe significant performance improvements across all five datasets. These results underline the effectiveness of SLGAN in NER tasks for both Chinese and English. Unlike Japanese, English is linguistically distinct from Chinese, further validating SLGAN's versatile effectiveness across different languages.

4.3.6 Results of Chinese Part-of-Speech Tagging Task

Table 6 provides a comprehensive overview of the evaluation results in the Chinese PoS Tagging task. Notably, SLGAN achieves the SOTA result on the CTB9 dataset, with an impressive improvement of up to **+1.46%** in the F1 score. It is worth mentioning that SLGAN, which relies solely on the semantic features of the text, outperforms the Glyce model (Meng et al., 2019) that leverages additional glyph information from Chinese characters. On the CTB5, CTB6, and UD1 datasets, BERT+DSC (Li et al., 2020) achieves SOTA results by replacing the conventional cross-entropy loss with dice loss, which effectively addresses data imbalance issues. To ensure a fair comparison, we contrast SLGAN's performance with BERT_CRF, which does not involve any loss function optimization. These results illustrate the competitiveness of SLGAN in PoS tagging tasks. SLGAN consistently outperforms BERT_CRF on all four datasets. Notably, there is a substantial improvement of up to **+2.93%** in the F1 score observed on the CTB9 dataset. Compared with LEBERT (Liu et al., 2021), which deeply integrates lexicon features into BERT, SLGAN achieves its results solely from the text input without relying on external features. The strong performance in Chinese PoS Tagging tasks underscores the universality of SLGAN.



Input1: 法官对费西列说，他将被释放，可以马上离开
SLGAN: [法官] [对] [费西列] [说] [，] [他] [将] [被] [释放] [，] [可以] [马上] [离开]
BERT-CRF: [法官] [对] [费] [西] [列] [说] [，] [他] [将] [被] [释放] [，] [可以] [马上] [离开]
Input2: 坪林的主要产业为茶，闻名的文山包种茶更是台湾茶中出类拔萃之品
SLGAN: [B-GPE] [E-GPE] [O] [O] [O] [O] [O] [O] [O] [O] [O] [O] [O] [O] [O] [O] ... [O]
BERT-CRF: [B-PER] [E-PER] [O] [O] [O] [O] [O] [O] [O] [O] [O] [O] [O] [O] [O] [O] ... [O]

Figure 2: A comparison between SLGAN and the output of BERT_CRF model.

4.3.7 Ablation Study and Qualitative Analysis

To further analyze the impact of GAN-generated pseudo-data on the performance of the sequence labeling model, we conducted an ablation experiment. We compare the initial output of the generator (traditional sequence labeling model) with the output after training (GAN-generated pseudo-data) to illustrate the model's learning process. We perform a case study by randomly selecting two examples from the CWS and Chinese NER tasks. In Figure 2, we observe that SLGAN aligns well with the ground truth in both samples. In the upper example, "The judge told Fesille that he would be released and could leave immediately", the main issue is with the entity "Fesille". In the lower sentence, "Tea is the primary industry in Pinglin, and the renowned Wenshan Pouchong tea stands out as one of the finest varieties of Taiwanese tea", the primary concern is with the entity "Pinglin". "Fesille" represents a person's name, which often poses a challenge in handling out-of-vocabulary (OOV) words. "Pinglin" typically relates to entities associated with administrative regions, and the entity type "GPE" usually signifies political entities like cities, states, countries, and continents, whereas "PER" represents a person's name. SLGAN correctly identifies entity boundaries but may misclassify entity types. These cases illustrate SLGAN's generalization ability in addressing sample bias.

5 Conclusion

SLGAN demonstrates its effectiveness in addressing sequence labeling tasks by leveraging GAN architecture. It offers a universal and efficient solution for tasks such as WS, NER, and PoS Tagging. SLGAN's performance is evaluated across six tasks spanning four languages, consistently achieving SOTA or near SOTA results in all experiments. Notably, SLGAN excels in handling OOV words, showcasing its SOTA performance in terms of R_oov. Furthermore, SLGAN extends its applicability to cross-domain CWS tasks. Moreover, SLGAN's impressive performance in NER and PoS Tagging tasks reiterates its versatility across different labeling tasks, transcending language barriers.

6 Limitations

Since SLGAN is built on the generative adversarial mechanism, it suffers from the inherent limitations of GAN networks and exhibits instability during training. Moreover, the training time will be longer than that of a framework composed of a single module. Additionally, the SLGAN framework currently focuses only on WS, NER, and PoS tagging tasks.

7 Acknowledgments

We sincerely appreciate and thank all the reviewers for their valuable feedback. This research was supported by the Scientific Research Starting Foundation of Shantou University under Project No.NTF23003.

References

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 2180–2188.

Key-Sun Choi, Young S Han, Young G Han, and Oh W Kwon. 1994. Kaist tree bank project for korean: Present and future development. In *Proceedings of the International Workshop on Sharable Natural Language Resources*, pages 7–14. Citeseer.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Ning Ding, Dingkun Long, Guangwei Xu, Muhua Zhu, Pengjun Xie, Xiaobin Wang, and Haitao Zheng. 2020. Coupling distant annotation and adversarial training for cross-domain Chinese word segmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6662–6671, Online. Association for Computational Linguistics.

Jingjing Gong, Xinchi Chen, Tao Gui, and Xipeng Qiu. 2019. Switch-lstms for multi-criteria chinese word segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6457–6464.

Hongyu Guo. 2015. Generating text with deep reinforcement learning. *arXiv preprint arXiv:1510.09202*.

Shohei Higashiyama, Masao Utiyama, Eiichiro Sumita, Masao Ideuchi, Yoshiaki Oida, Yohei Sakamoto, and Isaac Okada. 2019. Incorporating word attention into character-based word segmentation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2699–2709.

Weipeng Huang, Xingyi Cheng, Kunlong Chen, Taifeng Wang, and Wei Chu. 2020. Towards fast and accurate neural chinese word segmentation with multi-criteria learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2062–2072.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Yoshiaki Kitagawa and Mamoru Komachi. 2018. Long short-term memory for japanese word segmentation. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*.

Matt J Kusner and José Miguel Hernández-Lobato. 2016. Gans for sequences of discrete elements with the gumbel-softmax distribution. *arXiv preprint arXiv:1611.04051*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Alex M Lamb, Anirudh Goyal Alias Parth Goyal, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. 2016. Professor forcing: A new algorithm for training recurrent networks. In *Advances in neural information processing systems*, pages 4601–4609.

Gina-Anne Levow. 2006. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117, Sydney, Australia. Association for Computational Linguistics.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced nlp tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476.

Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. 2021. Lexicon enhanced Chinese sequence labeling using BERT adapter. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5847–5858, Online. Association for Computational Linguistics.

Yijia Liu, Yue Zhang, Wanxiang Che, Ting Liu, and Fan Wu. 2014. Domain adaptation for crf-based chinese word segmentation using free annotations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 864–874.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written japanese. *Language resources and evaluation*, 48(2):345–371.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, and Jiwei Li. 2019. Glyce: Glyph-vectors for chinese character representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 2746–2757. Curran Associates, Inc.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA. Association for Computational Linguistics.

Likun Qiu and Yue Zhang. 2015. Word segmentation for chinese novels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Alec Radford and Luke Metz. 2021. Unsupervised representation learning with deep convolutional generative adversarial networks.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242.

Yuanhe Tian, Yan Song, Fei Xia, Tong Zhang, and Yonggang Wang. 2020. Improving chinese word segmentation with wordhood memory networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8274–8285.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Yu Tong, Jingzhi Guo, and Jizhe Zhou. 2022a. Separation inference: A unified framework for word segmentation in east asian languages. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1521–1530.

Yu Tong, Jingzhi Guo, Jizhe Zhou, Ge Chen, and Guokai Zheng. 2022b. Word segmentation by separation inference for East Asian languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3924–3934, Dublin, Ireland. Association for Computational Linguistics.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807.

Shuang Wu, Xiaoning Song, and Zhenhua Feng. 2021. Mect: Multi-metadata embedding based cross-transformer for chinese named entity recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1529–1539.

Limao Xiong, Jie Zhou, Qunxi Zhu, Xiao Wang, Yuanbin Wu, Qi Zhang, Tao Gui, Xuan-Jing Huang, Jin Ma, and Ying Shan. 2023. A confidence-based partial label learning model for crowd-annotated named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1375–1386.

Zhenyu Xuan, Rui Bao, and Shengyi Jiang. 2020. Fgn: Fusion glyph network for chinese named entity recognition. In *China Conference on Knowledge Graph and Semantic Computing*, pages 28–40. Springer.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207–238.

Jiuding Yang, Jinwen Luo, Weidong Guo, Di Niu, and Yu Xu. 2023. Exploiting hierarchically structured categories in fine-grained chinese named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3407–3421.

Yuxiao Ye, Weikang Li, Yue Zhang, Likun Qiu, and Jian Sun. 2019. Improving cross-domain Chinese word segmentation with word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2726–2735, Minneapolis, Minnesota. Association for Computational Linguistics.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Qi Zhang, Xiaoyu Liu, and Jinlan Fu. 2018. Neural networks incorporating dictionaries for chinese word segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Yizhe Zhang, Zhe Gan, and Lawrence Carin. 2016. Generating text via adversarial training. In *NIPS workshop on Adversarial Training*, volume 21.

Yue Zhang and Jie Yang. 2018. Chinese NER using lattice LSTM. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1554–1564, Melbourne, Australia. Association for Computational Linguistics.

Enwei Zhu and Jinpeng Li. 2022. Boundary smoothing for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7096–7108.