Supplementary Material

Yu Tong TONGYU@STU.EDU.CN
Shantou University

Editors: Hung-yi Lee and Tongliang Liu

1. Validation on Biomedical Domain Datasets

To further validate the generalizability and practical applicability of the proposed method, we extended it to other low-resource or highly domain-specific NER tasks, such as those in the biomedical field. Specifically, we conducted experiments on a biomedical NER dataset and compared our approach with the method proposed in (Zhang et al., 2022). The experimental results demonstrate that our method also performs well in this domain, further confirming its strong cross-domain transferability and practical relevance. Compared to existing methods, our approach achieved a significant improvement in F1 score, highlighting its robustness in environments with scarce domain-specific terminology.

Model	CMeEE
BERT-base (Kenton and Toutanova, 2019)	62.1
BERT-wwm-ext-base (Cui et al., 2021)	61.7
RoBERTa-large (Liu et al., 2019b)	62.1
RoBERTa-wwm-ext-base	62.4
RoBERTa-wwm-ext-large	61.8
ALBERT-tiny (Lan et al., 2020)	50.5
ALBERT-xxlarge (Lan et al., 2020)	61.8
ZEN (Diao et al., 2020)	61.0
MacBERT-base (Cui et al., 2020)	60.7
MacBERT-large (Cui et al., 2020)	62.4
LPADA	**62.7**

Table 1: Evaluation of LPADA on Biomedical NER Datasets Against Existing Primary Methods.

2. Theoretical Analysis

In adversarial domain adaptation, the objective is to reduce the discrepancy between the source domain $\mathcal{D}_s$ and the target domain $\mathcal{D}_t$. When label information is incorporated, this problem can be framed as minimizing the difference between the **joint distributions** $P_s(X, Y)$ and $P_t(X, Y)$. To this end, we introduce the **Wasserstein distance** as a principled metric for measuring the discrepancy between these joint distributions:

$$W\left(P_s(X,Y), P_t(X,Y)\right) = \inf_{\gamma \in \Pi(P_s, P_t)} \mathbb{E}_{(x,y),(x',y') \sim \gamma} \left[\|(x,y) - (x',y')\| \right] \tag{1}$$

where:

- $\Pi(P_s, P_t)$ denotes the set of all joint distributions with marginals P_s and P_t,
- $\|\cdot\|$ is a distance metric.

This objective can be approximately optimized using an adversarial learning framework. To model the joint distribution, we concatenate the feature representation $f(x)$ with the corresponding label or prediction y, forming a joint feature representation $z = (f(x), y)$. A discriminator D is then trained to maximize the Wasserstein distance between the source and target joint representations, while the feature extractor G tries to minimize it:

$$\min_{G} \max_{D \in \mathcal{D}_L} \mathbb{E}_{(x_s, y_s) \sim \mathcal{D}_s} \left[D(f(x_s), y_s) \right] - \mathbb{E}_{x_t \sim \mathcal{D}_t} \left[D(f(x_t), \hat{y}_t) \right] \tag{2}$$

where:

- $\hat{y}_t$ denotes pseudo-labels assigned to target domain samples,
- $\mathcal{D}_L$ is the set of 1-Lipschitz functions.

Compared to aligning only the marginal feature distribution $P(X)$, aligning the joint distribution $P(X, Y)$ using Wasserstein distance provides several benefits:

- Stable training compared to other divergence measures like Jensen-Shannon (JS) divergence;
- Semantic consistency, as both feature and label information are considered;
- A principled measure of distribution alignment that mitigates issues such as mode collapse.

2.1. Motivation for Using Wasserstein Distance

Compared to conventional divergence metrics such as Jensen–Shannon (JS) (Lin, 2002) or Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951), the Wasserstein distance (also known as Earth Mover's Distance) offers the following theoretical advantages:

- **Well-defined**: It remains meaningful even when the supports of the two distributions do not overlap.
- **Continuity**: It is continuous over the space of probability measures, facilitating more stable optimization.
- **Geometric interpretability**: It measures the minimal cost of transforming one distribution into another, providing a more intuitive notion of discrepancy.

Therefore, Wasserstein distance is particularly suitable for aligning the joint distributions $P_s(X, Y)$ and $P_t(X, Y)$ in adversarial domain adaptation.

2.2. Joint Distribution Decomposition

The joint distribution can be factorized as:

$$P(X, Y) = P(Y|X)P(X) \tag{3}$$

Thus, aligning the joint distribution:

$$P_s(X, Y) \approx P_t(X, Y) \tag{4}$$

implicitly ensures the alignment of both the marginal feature distributions and the conditional label distributions:

$$P_s(X) \approx P_t(X), \quad P_s(Y|X) \approx P_t(Y|X) \tag{5}$$

This joint alignment strategy helps preserve both domain-invariant features and semantic consistency across domains.

2.3. Generalization Bound Perspective

Based on domain adaptation theory (Ben-David et al., 2010), the target error can be bounded by:

$$\epsilon_t(h) \leq \epsilon_s(h) + d_{\mathcal{H}\Delta\mathcal{H}}(P_s(X), P_t(X)) + \lambda \tag{6}$$

where:

- $\epsilon_s(h)$ is the source domain error,
- $d_{\mathcal{H}\Delta\mathcal{H}}$ measures the domain discrepancy (which can be approximated by the Wasserstein distance),
- λ is the minimum joint error of the ideal hypothesis on both domains.

By aligning $P(X, Y)$ instead of only $P(X)$, we effectively reduce both the domain discrepancy and the joint error term λ, thereby improving generalization on the target domain.

In summary, by aligning the joint distributions $P_s(X, Y)$ and $P_t(X, Y)$ using the Wasserstein distance, we simultaneously reduce both marginal and conditional shifts, providing a theoretically grounded and empirically stable approach for domain adaptation.

3. Rationale for Choosing the GAN-Based Framework

We acknowledge the potential of contrastive learning and natural language inference (NLI) in domain alignment tasks, particularly in terms of training stability, and agree that they are directions worth exploring. In this study, we chose to adopt GANs due to their capability to model complex data distributions under unsupervised or weakly supervised settings, which is especially suitable for our task where there exists a significant distributional shift between the source and target domains. Furthermore, previous studies have demonstrated that GANs can achieve higher-quality feature alignment in certain scenarios (e.g., DANN (Ganin et al., 2016)), which supports the suitability of our choice.

4. Implementation Details

4.1. Training Details and Optimizer Settings

The generator and discriminator were trained alternately for 10 epochs each. We used the Adam optimizer and applied an early stopping strategy to prevent overfitting.

4.2. Loss Function Weights

We adopted a hybrid loss, where the weighting ratio between the supervised loss and the adversarial loss was set to $\lambda = 1$. In preliminary experiments, we tested $\lambda \in \{0.5, 1, 2\}$, and found that $\lambda = 1$ yielded the most stable and effective performance. Therefore, this value was used in the final experiments.

4.3. Reproducibility

To ensure reproducibility, we will release the full training scripts along with the codebase, enabling researchers to replicate and further explore our method.

5. Details of the Compared Baseline Models

To ensure a comprehensive and fair evaluation of our proposed method, we select a diverse set of baseline models that encompass both traditional and state-of-the-art approaches in Named Entity Recognition (NER) and domain adaptation. Specifically: **(1)** BiLSTM-CRF serves as a classic and widely adopted sequence labeling framework, providing a strong traditional benchmark that does not rely on pre-trained contextual embeddings. It allows us to evaluate performance under a purely feature-driven architecture; **(2)** BERT-Chinese and RoBERTa are powerful pre-trained language models that offer rich contextualized representations and have demonstrated strong performance across a broad range of Chinese NLP tasks. Their inclusion enables us to assess how well general-purpose PLMs can adapt to the characteristics of our dataset. **(3)** SLGAN employs generative adversarial mechanisms to enhance sequence labeling performance, particularly under conditions of data scarcity or domain shift. Comparing against SLGAN provides insight into the effectiveness of adversarial training strategies relative to our approach. **(4)** CrossNER is specifically designed to evaluate cross-domain NER, with a strong focus on handling semantic discrepancies and label distribution mismatches across domains. Its inclusion is particularly relevant given the domain variability present in our task, offering a competitive benchmark for assessing generalization capabilities.

Together, these baselines cover a broad spectrum of methodological paradigms—including conventional sequence models, pre-trained language models, adversarial learning, and domain-adaptive architectures. This diverse baseline suite allows us to position our approach within the landscape of existing techniques and validate its effectiveness under both in-domain and cross-domain settings. We present below a detailed overview of the baseline models employed in our study.

- BiLSTM-CRF: we evaluate the Bidirectional LSTM with CRF decoding (Huang et al., 2015) as a foundational reference model. As one of the most classic and widely adopted architectures for Named Entity Recognition (NER), BiLSTM-CRF effectively captures contextual dependencies through bidirectional sequence modeling, while the CRF layer enforces valid label transitions and globally optimized predictions. Its inclusion serves to establish a strong

traditional benchmark, facilitating meaningful comparison with more advanced or domain-adaptive methods in our experiments.

- BERT: we adopt BERT-Chinese (Devlin et al., 2019) as a representative deep learning baseline. As a pre-trained language model widely used in various Chinese NLP tasks, it provides a strong and consistent reference point for evaluating the effectiveness of more advanced or task-specific approaches on our dataset. Its inclusion ensures that performance improvements can be assessed relative to a well-established benchmark.

- RoBERTa: to enhance the comprehensiveness of our evaluation, we further incorporate RoBERTa (Liu et al., 2019a), a robustly optimized pre-trained language model that builds upon BERT with improved training strategies, including dynamic masking, larger mini-batches, and longer training durations. RoBERTa has demonstrated superior performance across a wide range of NLP benchmarks and serves in our study as a strong high-capacity baseline. By leveraging its rich contextual representations, we are able to assess the upper-bound performance achievable by general-purpose pre-trained models on our dataset. This comparison also helps to highlight the potential benefits of introducing domain-specific enhancements or task-oriented adaptations.

- SLGAN: to provide a more recent and competitive benchmark for the TCMNER2024 dataset, we include SLGAN (Tong et al., 2024), a model that leverages generative adversarial learning to enhance sequence labeling performance. By introducing adversarial training dynamics, SLGAN encourages the feature extractor to generate more robust and domain-invariant representations, thereby improving generalization across entity types and linguistic variations. Its inclusion enables us to compare our approach against state-of-the-art generative frameworks specifically designed for NER tasks.

- CrossNER: By incorporating CrossNER (Jia et al., 2019) into our study, we aim to assess how well domain-adaptive models can handle the domain gap present in the TCMNER2024 dataset. This inclusion also allows us to examine the limitations of existing approaches under real-world transfer scenarios and further motivates the need for more robust generalization techniques.

6. Annotation Details

To ensure consistency and accuracy throughout the annotation process, we developed specialized annotation tools along with detailed usage instructions. Annotators underwent thorough training and calibration to guarantee a comprehensive understanding of the guidelines and their consistent application.

6.1. Data Collection

We collected over 5,000 clinical case documents from more than 400 distinguished TCM physicians, sourced from a diverse range of published materials and authoritative texts. Our dataset, TCMNER2024, is publicly available on GitHub [1]. To protect sensitive information, we utilized only openly accessible records and ensured full anonymization throughout the dataset. Text processing

1. https://github.com/TCMNER/TCMNER2024

centered on segments describing patients' first visits, and annotations were carried out based on the presented symptoms. The rationale for focusing on initial consultations is outlined as follows:

- To limit the text length, as including the second and third consultations would make the text too long.
- To maintain a higher density of entity words, since the second and third consultations contain fewer annotatable sections.
- To avoid the less formal and less standardized language often found in the texts of the second and third consultations.

6.2. Data Preprocessing

Text Cleaning

- Noise Removal: Unusable content such as encoding distortions, malformed symbols, redundant records, and lines lacking labeled entities was discarded.
- Content Filtering: Only samples rich in domain-relevant elements—such as TCM concepts, clinical manifestations, medicinal formulations, and therapeutic strategies—were preserved to maintain informational value.
- Sentence Validation: Each text unit was checked for semantic coherence to eliminate incomplete expressions, especially those introduced by faulty segmentation or line disruptions in original medical sources.

Denoising

- Filtering out unreadable elements, including distorted symbols, encoding glitches, and extraneous non-standard marks.
- Cleaning structural clutter like excessive punctuation, irregular spacing, and residual tags from digitized text sources.
- Discarding near-identical entries to avoid bias caused by repetition and overly similar training instances.

6.3. Annotation Guideline

- Annotation should capture only the essential meaning of the entity, omitting non-essential modifiers and auxiliary terms. For example:
  Correct: "脾气虚" (Spleen Qi Deficiency) should be annotated as the entity. Incorrect: "伴有明显的脾气虚表现" (accompanied by obvious signs of spleen qi deficiency) → Do not include "伴有明显的" (accompanied by obvious) or "表现" (signs) in the annotation.
- Entities expressing the same concept are labeled in a uniform manner across all documents. A curated vocabulary list is provided to ensure annotation standardization.

- No Entity Nesting: Nested entity structures are prohibited. An illustrative example is provided below for clarity.
  Nested structure (e.g., "桂枝汤合四逆汤" (Combination of Guizhi Decoction and Sini Decoction)) is annotated by selecting only the core entities (e.g., "桂枝汤" (Guizhi Decoction) and "四逆汤" (Sini Decoction) as separate formulas); connective terms such as "合" (combination of) are not included in the annotation.

- Annotation at the Phrase Level
  Entities are annotated at the phrase level rather than word by word. For example, "口干咽燥" (dry mouth and throat) is annotated as a complete symptom, rather than tagging individual characters separately.

6.4. Quality Control

Each data instance is independently labeled by two annotators, followed by a third-party review to ensure annotation reliability and consistency. To support knowledge sharing and resolve edge cases, a centralized repository of ambiguous or challenging examples is maintained and regularly updated with adjudication outcomes and guideline clarifications. Inter-annotator agreement (IAA) metrics, such as Cohen's Kappa or F1-based measures, are periodically computed to assess annotator consistency, with targeted feedback or retraining provided when necessary.

A comprehensive annotation guideline and a quick-reference cheatsheet are provided to all annotators to ensure task clarity and labeling standardization. Annotators must pass a qualification test before participating in the actual annotation, and periodic spot-checks are conducted to monitor annotation accuracy. Only annotations with full agreement between the two annotators are accepted; in cases of disagreement, the instance is escalated for expert adjudication. Additionally, a structured feedback mechanism allows annotators to flag uncertain cases, ambiguous definitions, or potentially missing entity types. These cases are reviewed by domain experts, whose decisions are logged and reflected in iterative updates to the annotation guidelines. All revisions to the guidelines are version-controlled and communicated to annotators through briefing sessions. This expert-driven adjudication and continuous feedback loop ensures high-quality, consistent, and reproducible annotations throughout the dataset.

7. Ablation Study

To evaluate the individual contributions of key components—including adversarial training, the attention mechanism, the sequence tagging module, and the metric used for measuring distribution alignment—we perform a series of ablation studies under the TCM NER task setting. The experiments are designed as follows:

- Removal of adversarial training: to assess its role in mitigating domain shift.

- Exclusion of the attention mechanism: to examine its impact on entity boundary recognition.

- Replacement of the sequence tagger: to test how different decoding architectures affect performance.

- Substituting the Wasserstein distance with a naive alternative (e.g., L1 distance): to test whether the proposed Wasserstein-based distribution alignment contributes significantly beyond simple distance-based losses.

7.1. Impact of Adversarial Training

We compare the model without adversarial training (using only the generator part, i.e., applying the general sequence labeling algorithm BERT+CRF, trained on the target domain training set and evaluated on the test set). Table 2 presents the comparison results, which demonstrate that without adversarial training-i.e., without domain adaptation, the model performs poorly when limited data is available in the low-resource domain. Removing adversarial training leads to a consistent drop of 0.91% in F1 across target domains, indicating its importance in reducing distributional shifts. Therefore, adversarial domain adaptation plays a crucial role in the overall framework.

Entity	No Adversarial	LPADA
TCM Disease Terms	41.04	**42.33**
Western Medicine Disease Terms	74.23	**75.66**
TCM Symptoms	49.07	**50.88**
Urination and Defecation	80.75	**81.86**
Pulse Conditions	95.36	**96.10**
Tongue Conditions	94.66	**95.39**
Western Medicine Symptoms	47.18	**48.41**
TCM Syndromes	85.01	**85.48**
TCM Therapeutic Principle	88.97	**89.27**
TCM Prescriptions	78.68	**79.42**
Chinese Materia	97.43	**97.61**
Average	75.67	**76.58**

Table 2: Comparison between adversarial training and non-adversarial training, with F1 score as the evaluation metric.

Entity	No Attention	LPADA
TCM Disease Terms	40.85	**42.33**
Western Medicine Disease Terms	75.02	**75.66**
TCM Symptoms	49.43	**50.88**
Urination and Defecation	80.87	**81.86**
Pulse Conditions	95.45	**96.10**
Tongue Conditions	94.53	**95.39**
Western Medicine Symptoms	47.24	**48.41**
TCM Syndromes	84.63	**85.48**
TCM Therapeutic Principle	88.15	**89.27**
TCM Prescriptions	78.52	**79.42**
Chinese Materia	97.20	**97.61**
Average	75.63	**76.58**

Table 3: Comparison between models with and without the attention mechanism, with F1 score as the evaluation metric.

7.2. Impact of Attention Mechanism

Table 3 presents the performance of our proposed model after removing the attention mechanism. Excluding the attention mechanism results in a 0.95% decrease in precision, suggesting that attention enhances entity boundary detection. In this variant, the model reduces to a framework that focuses solely on aligning source and target domain feature distributions, but lacks the capacity for dynamic label-aware representation learning. Specifically, the absence of the attention mechanism hinders the model's ability to selectively attend to label-relevant contextual cues. This limitation leads to a more shallow and coarse-grained alignment, where the model struggles to distinguish subtle differences across entity types or adapt to target-domain label semantics. As reflected in the results, this simplification results in a noticeable degradation in overall performance, particularly in the recognition of less frequent or domain-specific entity types. These findings highlight the essential role of the attention mechanism not only in enhancing context-sensitive encoding but also in enabling fine-grained, label-specific adaptation, which is crucial for robust cross-domain named entity recognition.

7.3. Impact of Sequence Tagger

To verify the impact of CRF on the overall framework, we replace it with the widely used softmax as the label classifier. Table 4 compares different sequence taggers, and the results show that the choice of sequence tagger has a negligible impact on the final experimental results.

Entity	Softmax	CRF
TCM Disease Terms	**42.35**	42.33
Western Medicine Disease Terms	75.63	**75.66**
TCM Symptoms	50.85	**50.88**
Urination and Defecation	**81.87**	81.86
Pulse Conditions	96.08	**96.10**
Tongue Conditions	95.37	**95.39**
Western Medicine Symptoms	**48.45**	48.41
TCM Syndromes	85.45	**85.48**
TCM Therapeutic Principle	**89.30**	89.27
TCM Prescriptions	79.39	**79.42**
Chinese Materia	97.58	**97.61**
Average	76.57	**76.58**

Table 4: Comparison between softmax and CRF, with F1 score as the evaluation metric.

7.4. Impact of Metrics for Measuring Data Distribution

In the discriminator, we use Wasserstein loss to measure the distribution discrepancy between the source and target domain data. To assess the impact of this component, we replace it with alternative metrics for measuring distribution differences, such as L1 distance. Table 6 presents the comparison results, which show that different metrics for measuring data distribution have a negligible effect on the overall framework.

In conclusion, as shown in Table 5, the performance gains of the proposed LPADA framework in NER tasks primarily stem from the incorporation of domain-adaptive adversarial training and the

Model Variant	Adv. Training	Attention	Sequence Tagger	Wasserstein Loss	Average F1	Drop from Full
Full Model	✓	✓	✓	✓	76.58	—
No Adv. Training		✓	✓	✓	75.67	↓0.91
No Attention	✓		✓	✓	75.63	↓0.95
Softmax (Sequence Tagger)	✓	✓		✓	76.57	↓0.01
L1 Distance (Metric)	✓	✓	✓		76.57	↓0.01

Table 5: Quantitative Impact of Individual Components.

Entity	L1 Distance	Wasserstein Loss
TCM Disease Terms	42.31	**42.33**
Western Medicine Disease Terms	**75.68**	75.66
TCM Symptoms	50.84	**50.88**
Urination and Defecation	81.83	**81.86**
Pulse Conditions	**96.12**	96.10
Tongue Conditions	**95.40**	95.39
Western Medicine Symptoms	48.40	**48.41**
TCM Syndromes	85.46	**85.48**
TCM Therapeutic Principle	**89.28**	89.27
TCM Prescriptions	79.40	**79.42**
Chinese Materia	97.58	97.61
Average	76.57	**76.58**

Table 6: Comparison between L1 distance and Wasserstein Loss, with F1 score as the evaluation metric.

attention mechanism. In contrast, variations in the choice of distribution discrepancy metrics and sequence tagging modules appear to have relatively minor effects on the overall effectiveness of the framework. This finer-grained analysis strengthens the empirical rigor of our study and clarifies how each component contributes to overall performance.

Figure 1: A comparison between LPADA and the traditional BERT_CRF model.

8. Qualitative Analysis

We compare the generator's output (traditional sequence labeling model) with the result after adversarial training to highlight the model's learning progression. We perform a case study by randomly selecting two examples from the TCM NER task. In Figure 1, we observe that LPADA aligns well with the ground truth in both samples. In the upper example, *"The patient experiences damp-heat in both feet, with swelling and burning pain that gradually spreads to the groin. Recurrent JiaoQi (脚气) also occur. Jiawei Ermiaowan can be used for treatment."* In the field of Western medicine or in a broader medical context, JiaoQi("脚气") usually refers to a foot condition caused

by fungal infection, it is commonly translated as "athlete's foot" or "dermatophytosis". However, In the specific field of Traditional Chinese Medicine (TCM), "JiaoQi(脚气)", which could be translate as "foot dampness", is a unique TCM disease term for a condition characterized primarily by numbness, soreness, weakness, cramping, swelling, or atrophy of the legs and feet. The traditional BERT+CRF model mistakenly recognized it as a Western medical disease. In the lower sentence, *"Due to the patient's blood stasis and poor circulation ("血瘀不行"), myrrh and frankincense are added to the formula to invigorate blood circulation and resolve stasis."* "Stasis and poor circulation(血瘀不行)" is a specialized term categorized as a "syndrome" in TCM, which should be recognized and understood as a whole. In contrast to BERT+CRF, LPADA demonstrates superior accuracy in identifying entity boundaries and types. These examples highlight LPADA's ability to generalize and mitigate sample bias.

References

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1–2):151–175, May 2010. ISSN 0885-6125. doi: 10.1007/s10994-009-5152-4. URL https://doi.org/10.1007/s10994-009-5152-4.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for Chinese natural language processing. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.58. URL https://aclanthology.org/2020.findings-emnlp.58/.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. Pre-training with whole word masking for chinese bert. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3504–3514, November 2021. ISSN 2329-9290. doi: 10.1109/TASLP.2021.3124365. URL https://doi.org/10.1109/TASLP.2021.3124365.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. Zen: Pre-training chinese text encoder enhanced by n-gram representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4729–4740, 2020.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

Chen Jia, Xiaobo Liang, and Yue Zhang. Cross-domain ner using cross-domain language modeling. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2464–2474, 2019.

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

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2020. URL https://arxiv.org/abs/1909.11942.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 2002.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019a.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019b. URL https://arxiv.org/abs/1907.11692.

Yu Tong, Ge Chen, Guokai Zheng, Rui Li, and Jiang Dazhi. When generative adversarial networks meet sequence labeling challenges. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10625–10635, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.593. URL https://aclanthology.org/2024.emnlp-main.593/.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.544. URL https://aclanthology.org/2022.acl-long.544/.