Federated Chinese Word Segmentation with Global Character Associations

Yuanhe Tian[♥*], **Guimin Chen**[◇*], **Han Qin**[♠*], **Yan Song**[♠♡†]
[♥]University of Washington [◇]QTrade
[♠]The Chinese University of Hong Kong (Shenzhen)
[♡]Shenzhen Research Institute of Big Data
[♥]yhtian@uw.edu [◇]chenguimin@foxmail.com
[♠]hanqin@link.cuhk.edu.cn [♠]songyan@cuhk.edu.cn

Abstract

Chinese word segmentation (CWS) is a fundamental task for Chinese information processing, which always suffers from out-of-vocabulary word issues, especially when it is tested on data from different sources. Although one possible solution is to use more training data, in real applications, these data are stored at different locations and thus are invisible and isolated among each other owing to the privacy or legal issues (e.g., clinical reports from different hospitals). To address this issue and benefit from extra data, we propose a neural model for CWS with federated learning (FL) adopted to help CWS deal with data isolation, where a mechanism of global character associations is proposed to enhance FL to learn from different data sources. Experimental results on a simulated environment with five nodes confirm the effectiveness of our approach, where our approach outperforms different baselines including some well-designed FL frameworks.[1]

1 Introduction

Chinese word segmentation (CWS) is a preliminary and vital task for natural language processing (NLP). This task aims to segment Chinese character sequence into words and thus is generally performed as a sequence labeling task (Tseng et al., 2005; Levow, 2006; Song et al., 2009a; Sun and Xu, 2011; Song and Xia, 2012, 2013; Mansur et al., 2013). Although recent neural-based CWS systems (Pei et al., 2014; Chen et al., 2017; Ma et al., 2018; Higashiyama et al., 2019; Qiu et al., 2019; Ke et al., 2020; Huang et al., 2020a; Tian et al., 2020e) have achieved very good performance on benchmark datasets, it is still an unsolved task (Fu et al., 2020), because it is challenging to handle out-of-vocabulary words (OOV), especially in real applications where the test data may come from different sources. Although leveraging extra labeled data from other sources or domains could alleviate this issue, in real applications, such data are always located in different nodes and thus are inaccessible to each other because of the privacy or legal concerns (e.g., clinical or financial reports from different hospitals or companies).

To address the data isolation issue, federated learning (FL) (Shokri and Shmatikov, 2015; Konečný et al., 2016) is proposed and has shown its great promises for many machine learning tasks (Aono et al., 2017; Sheller et al., 2018; He et al., 2020). In many cases, data in different nodes are encrypted and aggregated to the centralized model, and they are invisible to each other during the training stage. This property allows FL to be an essential technique for real applications with privacy and security requirements. However, conventional FL techniques are more suitable for nodes sharing homogeneous data, which is seldom the case for NLP tasks. Particularly for CWS, the appropriate segmentation is sensitive to the data source, where the text and vocabularies used in different datasets contain various expressing patterns. For example, in real applications such as Input Method Editors (IME, such as pinyin input environment), there are millions of individual users with their data stored in isolated nodes, where the different nodes could have diverse segmentation requirement due to the users' preference. Therefore, the restricted data access of traditional FL approaches could result in inferior performance for CWS since they cannot update the model to facilitate localized prediction. Unfortunately, limited attentions have been paid to address this issue. Most existing approaches (Liu et al., 2019; Huang et al., 2020b; Sui et al., 2020) with FL on NLP (e.g., for language modeling

[*] Equal contribution.

[†] Corresponding author.

[1] The code and models involved in this paper are released at https://github.com/cuhksz-nlp/GCASeg.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4306–4313
August 1–6, 2021. ©2021 Association for Computational Linguistics

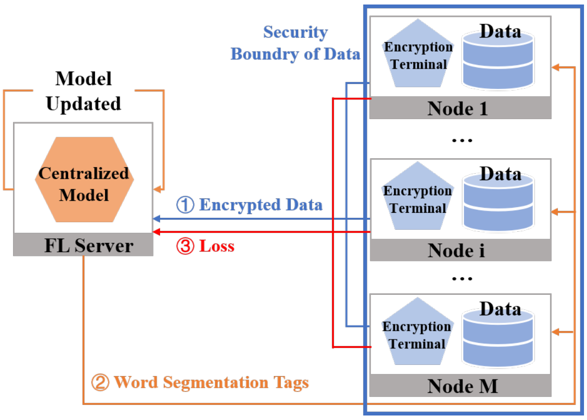

Figure 1: The server-node architecture of our approach. The encrypted information (i.e., encrypted data, word segmentation tags, and loss) communicates between a node and the server, where the locally stored data is inaccessible to other nodes during the training process.

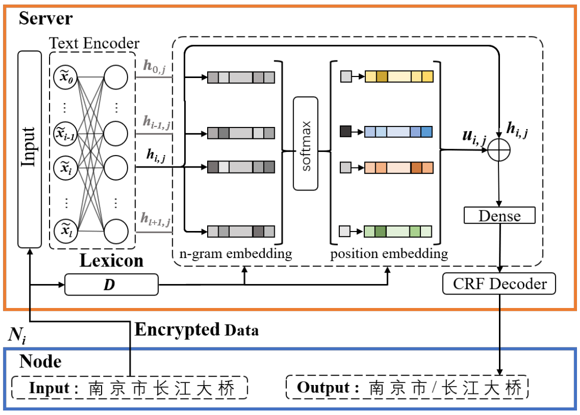

Figure 2: An overview of GCA-FL, where centralized model is illustrated on the top and the example input sentence "南京市长江大桥" (*Nanjing Yangtze River Bridge*) from $\mathcal{N}_i$ is shown on the bottom. "/" in the output represents the delimiter for the word boundaries.

(Hard et al., 2018; Chen et al., 2019), named entity recognition (Ge et al., 2020), and text classification (Zhu et al., 2020)) mainly focus on optimizing the learning process and ignore domain diversities.

In this paper, we propose a FL-based neural model (GCA-FL) for CWS, which is enhanced by global character association (GCA) mechanism in a distributed environment. The GCA mechanism is designed to capture contextual information (patterns) in a particular input for localized predictions and to handle the difficulties in identifying text sources caused by data inaccessibility. Specifically, GCA is served as a server-side component to associate global character n-grams with different inputs from each node and responds with contextual information to help the backbone segmenter. Experimental results on a simulated environment with isolated data from five domains demonstrate the effectiveness of our approach, where GCA-FL outperforms different baselines including the ones with well designed FL framework.

2 The approach

Figure 1 illustrates the overall server-node architecture for applying our approach. The centralized model is stored in the FL server and data from multiple sources (domains) are stored at different nodes (the i-th node is denoted by $\mathcal{N}_i$), respectively. Encrypted information (e.g., data, vectors, and loss) communicates between each node $\mathcal{N}_i$ and the FL server. In this way, the original data stay in the local node and is not accessible to the other nodes. To encode contextual information (patterns) to facilitate localized prediction, we enhance FL by introducing GCA into the centralized model (Figure 2), which follows the character-based sequence labeling paradigm for CWS. Herein, GCA encodes the contextual information from the encrypted input and uses the resulted information to guide the centralized model to make a localized prediction. In the following, we introduce FL for CWS and then the centralized model with GCA.

2.1 Federated Learning

In the training process of FL, the node $\mathcal{N}_i$ firstly encrypts the original input character sequence $\mathcal{X}_i$ into $\widetilde{\mathcal{X}}_i = \widetilde{x}_{i,1} \cdots \widetilde{x}_{i,j} \cdots \widetilde{x}_{i,l}$, where $\widetilde{x}_{i,j}$ denotes the j-th character in $\widetilde{\mathcal{X}}_i$. Next, $\mathcal{N}_i$ passes $\widetilde{\mathcal{X}}_i$ to the server. Then, the centralized model on the server processes $\widetilde{\mathcal{X}}_i$ and predicts the corresponding label sequence $\widehat{\mathcal{Y}}_i = \widehat{y}_{i,1} \cdots \widehat{y}_{i,j} \cdots \widehat{y}_{i,l}$ by

$$\widehat{\mathcal{Y}}_i = \text{GCA-FL}\left(\widetilde{\mathcal{X}}_i\right) \quad (1)$$

where $\widehat{y}_{i,j} \in \mathcal{T}$ ($\mathcal{T}$ is the label set) is the segmentation label for $\widetilde{x}_{i,j}$. Afterwards, $\widehat{\mathcal{Y}}_i$ is passed back to $\mathcal{N}_i$ and compared with the gold label sequence $\mathcal{Y}_i^*$, after which the loss $\mathcal{L}_i$ for that training instance is obtained locally. Finally, $\mathcal{L}_i$ is passed to the server and the parameters in the centralized model are updated accordingly.

2.2 Centralized Model with GCA

In standard FL framework, the backbone centralize model works following the encoding-decoding paradigm, where $\widetilde{\mathcal{X}}_i$ is encoded[2] into a sequence of hidden vectors ($\mathbf{h}_{i,j}$ denotes the hidden vector for $\widetilde{x}_{i,j}$), which are then sent to a decoder (e.g.,

[2] One can use any encoder, e.g., biLSTM, for this process.

Rule	$v_{i,j,k}$
$x_{i,j}$ is the beginning of the n-gram $s_{i,j,k}$	V_B
$x_{i,j}$ is inside the n-gram $s_{i,j,k}$	V_I
$x_{i,j}$ is the ending of the n-gram $s_{i,j,k}$	V_E
$x_{i,j}$ is the single-character n-gram $s_{i,j,k}$	V_S

Table 1: The rules for assigning different position patterns to $x_{i,j}$ based on its position in an n-gram $s_{i,j,k}$.

CRF) to obtain the prediction $\widehat{\mathcal{Y}}_i$. However, data from different nodes (sources) always contains heterogeneous vocabularies and expressing patterns, where standard FL may obtain inferior results for localized prediction because it cannot distinguish the contextual information from the isolated data. Therefore, motivated by previous studies that leverage n-grams to capture local contextual information (Song et al., 2009b; Pei et al., 2014; Chen et al., 2017; Higashiyama et al., 2019), we propose GCA to enhance standard FL by exploring the contextual information carried by n-grams in the running text and use it to guide the centralized model for making localized prediction.

Specifically, GCA contains three components, namely, a lexicon ($\mathcal{D}$) that contains global character n-grams, an n-gram embedding matrix that maps an n-gram in $\mathcal{D}$ to its embedding, and a position embedding matrix that maps a position pattern, i.e., the position (e.g., beginning, ending, and inside) of a character in an n-gram, to its embedding. For each character $\widetilde{x}_{i,j}$, GCA encodes the contextual information and uses it to enhance the centralized model in the following process. First, GCA extracts all m n-grams $s_{i,j,k}$ $(1 \leq k \leq m)$ associated with $\widetilde{x}_{i,j}$ from $\mathcal{D}$, where $s_{i,j,k}$ satisfies the conditions that it contains $\widetilde{x}_{i,j}$ and it is a sub-string of $\widetilde{\mathcal{X}}_i$. Next, according to the position of $\widetilde{x}_{i,j}$ in $s_{i,j,k}$, GCA finds the position pattern $v_{i,j,k}$ associated with $s_{i,j,k}$ and $\widetilde{x}_{i,j}$ based on the rules specified in Table 1. For example, if $\widetilde{x}_{i,j}$ =“市” (*city*) and $s_{i,j,k}$ =“市长” (*mayor*), the position pattern $v_{i,j,k}$ will be “V_B” according to the rules in Table 1, because $\widetilde{x}_{i,j}$ is at the beginning of $s_{i,j,k}$. Third, GCA applies n-gram embedding matrix and position embedding matrix to $s_{i,j,k}$ and $v_{i,j,k}$, respectively, and obtains the n-gram embedding $\mathbf{e}^s_{i,j,k}$ and the position embedding $\mathbf{e}^v_{i,j,k}$. Then, GCA computes the weights $p_{i,j,k}$ for position patterns $v_{i,j,k}$ by

$$p_{i,j,k} = \frac{\exp(\mathbf{h}_{i,j} \cdot \mathbf{W} \cdot \mathbf{e}^s_{i,j,k})}{\sum_{k=1}^{m} \exp(\mathbf{h}_{i,j} \cdot \mathbf{W} \cdot \mathbf{e}^s_{i,j,k})} \quad (2)$$

Genres		Sent. #	Token #	OOV Rate
BC	Train	7,301	628K	-
	Dev	2,368	237K	7.6
	Test	2,383	235K	6.1
BN	Train	5,867	1,108K	-
	Dev	2,151	394K	2.7
	Test	2,068	427K	4.2
MZ	Train	5,250	970K	-
	Dev	1,647	299K	2.4
	Test	1,526	342K	4.5
NW	Train	6,603	1,094K	-
	Dev	2,029	331K	4.1
	Test	2,085	346K	1.4
Web	Train	6,094	839K	-
	Dev	2,085	285K	4.2
	Test	2,001	243K	8.2

Table 2: The number of sentences, word tokens, and the out-of-vocabulary (OOV) rate (in dev and test sets) with respect to the training set of five genres in CTB7.

where $\mathbf{W}$ is a trainable matrix that maps $\mathbf{e}^s_{i,j,k}$ to the same dimension as $\mathbf{h}_{i,j}$ to facilitate the inner production “$\cdot$”. GCA further applies $p_{i,j,k}$ to all position embeddings and obtain the representation of contextual information $\mathbf{u}_{i,j}$ for $\widetilde{x}_{i,j}$ by

$$\mathbf{u}_{i,j} = \sum_{k=1}^{m} p_{i,j,k} \cdot \mathbf{e}^v_{i,j,k} \quad (3)$$

Afterwards, $\mathbf{u}_{i,j}$ is added (+) to $\mathbf{h}_{i,j}$ to guide the backbone model for localized prediction, where the resulted vector is mapped into the output space by a trainable matrix $\mathbf{W}_o$ and bias $\mathbf{b}_o$ by

$$\mathbf{o}_{i,j} = \mathbf{W}_o \cdot (\mathbf{u}_{i,j} + \mathbf{h}_{i,j}) + \mathbf{b}_o \quad (4)$$

Finally, $\mathbf{o}_{i,j}$ is fed into a CRF decoder to obtain the predicted segmentation label $\widehat{y}_{i,j}$ for $\widetilde{x}_{i,j}$.

3 Experimental Settings

3.1 Simulations

To test the proposed approach, we follow the convention of recent FL-based NLP studies (Liu et al., 2019; Huang et al., 2020b; Zhu et al., 2020; Sui et al., 2020) to build a simulated environment where isolated data are stored in five nodes. Each node contains one of the five genres (i.e., broadcast conversation (BC), broadcast news (BN), magazine (MZ), newswire (NW), and weblog (WEB)) in CTB7 (LDC2010T07)[3] for CWS. Therefore, data from different genres are distributed to the five nodes without overlapping (i.e., the data sources in

[3] https://catalog.ldc.upenn.edu/LDC2010T07

Hyper-parameters	Values
Learning Rate	$5e-6$, **1e-5**, $3e-5$
Warmup Rate	**0.1**, 0.2
Dropout Rate	**0.33**
Batch Size	**16**, 32

Table 3: The hyper-parameters tested in tuning our models. The best ones used in our final experiments are highlighted in boldface.

our simulation are heterogeneous), which is similar to the simulation setting of aforementioned previous studies. We split each genre into train/dev/test splits following Wang et al. (2011) and report the statistics (in terms of the number of sentences, word tokens, and OOV rate) in Table 2.

3.2 Baselines and Reference Models

To show the effectiveness of our approach with GCA, we compare it with a baseline model that follows the FL framework without using it. In addition, we also run two reference models without both FL and GCA, where all training instances are not isolated and are accessible to each other. Specifically, the first reference model (denoted by Single) is trained and evaluated on the data from a single node (genre). The second (denoted by Union) is trained on the union of training instances from all five nodes (genres) and evaluated on a single node. Herein, the Union reference model can be optimized on a particular local node to achieve the best localized prediction; on the contrary, models under the FL setting is stored on the server and shared by all nodes, so that optimizing the model on a particular node could significantly hurt the performance on others. Therefore, the setting of the Union reference model is the ideal situation which is hard to happen in real-applications and it thus provides a potential upper-boundary of model performance for FL-based approaches.

3.3 Implementation

A good text representation is generally a prerequisite to achieve outstanding model performance (Pennington et al., 2014; Song and Shi, 2018; Peters et al., 2018). To obtain a high qulity of text representation, in our experiments, we try two types of encoder in the centralized model, i.e., the Chinese version of BERT (Devlin et al., 2019)[4] and the large version of ZEN 2.0 (Song et al., 2021)[5], because they are pre-trained language models that have been demonstrated to be effective in many NLP tasks (Nie et al., 2020; Huang et al., 2020a; Song et al., 2020; Chen et al., 2020; Fu et al., 2020; Tian et al., 2020a,b,c,d, 2021a,b; Chen et al., 2021; Qin et al., 2021). For both BERT and ZEN 2.0, we use the default settings (i.e., 12 layers of multi-head attentions with 768 dimensional hidden vectors for BERT and 24 layers of multi-head attentions with 1024 dimensional hidden vectors for ZEN 2.0). We use the vocabulary in Tencent Embedding[6] (Song et al., 2018) to initialize our lexicon $\mathcal{D}$ and the n-gram embedding matrix, where n-grams whose character-based length higher than five are filtered out[7]. During the training stage, we fix the n-gram embedding matrix and update all other parameters (including BERT). For evaluation, we follow previous studies to use the F1 scores (Chen et al., 2017; Ma et al., 2018; Qiu et al., 2019). For other hyper-parameter settings, we report them in Table 3. We test all combinations of them for each model on the development set, where models achieve highest F1 score on the development set is evaluated on the test set (the best hyper-parameter setting in our experiments is highlighted in boldface).

4 Results and Analysis

4.1 Overall Results

Table 4 illustrates the experimental results (i.e., F1 scores) of our GCA-FL models and all the aforementioned baselines (i.e., FL) and reference models (i.e., Single and Union) with BERT (a) and ZEN 2.0 (b) encoders on the test set of BC, BN, MZ, NW, and Web from CTB7.

There are several observations from the test set results. First, models under the FL framework (i.e., FL and GCA-FL) outperform the reference model (Single) trained on the single node for both BERT and ZEN 2.0 encoder, which confirms that FL works well to leverage extra isolated data. Second, our GCA-FL model consistently outperforms the FL baseline on all nodes (genres), although the FL baseline with BERT and ZEN 2.0 has already achieved very good performance. This observation demonstrates the effectiveness of the proposed GCA mechanism to leverage contextual information to facilitate localized prediction. Third, it is

[4] We use the Chinese base model from `https://s3.amazonaws.com/models.huggingface.co/`.

[5] We download the Chinese version of ZEN 2.0 from `https://github.com/sinovation/ZEN2`.

[6] `https://ai.tencent.com/ailab/nlp/en/embedding.html`

[7] We use five as the threshold because most Chinese words contain no more than five characters.

	BC	BN	MZ	NW	WB	Avg.
Single	97.13	96.97	96.21	97.84	94.83	96.60
Union	97.80	97.49	96.74	98.44	95.30	97.15
FL	97.49	97.22	96.54	98.15	95.03	96.89
GCA-FL	**97.76**	**97.40**	**96.74**	**98.43**	**95.29**	**97.12**

(a) BERT

	BC	BN	MZ	NW	WB	Avg.
Single	97.43	97.38	96.33	98.11	95.14	96.88
Union	97.88	97.79	97.23	98.61	96.38	97.58
FL	97.62	97.56	96.88	98.44	95.88	97.28
GCA-FL	**97.83**	**97.76**	**97.01**	**98.50**	**95.94**	**97.41**

(b) ZEN 2.0

Table 4: Experimental results (i.e., F1 scores) of different models with BERT (a) and ZEN 2.0 (b) on the development sets of the five nodes (genres) of CTB7.

observed that GCA-FL achieves competitive results compared with the Union reference model in most cases. This observation is promising because the Union model has all training data available without suffering from the data isolation problem, which could provide a potential upper boundary for FL-based models. The results obtained from GCA-FL thus further confirm the effectiveness of GCA.

4.2 Effect of GCA

To analize the effect of GCA to leverage isolated extra data to facilitate localized prediction, especially for OOV, we illustrate the recall of OOV of different models (i.e., Single, Union, FL, and GCA-FL) on five nodes (genres) with BERT encoder in Figure 3. Similar to the experimental results in the main experiments, it is observed that FL and GCA-FL outperform Single model in identifying unseen words (OOV). Further, GCA-FL can outperform the FL baseline on the test data in all nodes, where the highest improvement is observed on the node storing data from newswire. One possible explanation could be that BC and BN contains similar texts to NW. GCA-FL can better learn from the similar data on these nodes and thus improves localized prediction, especially for OOV.

5 Conclusion

In this paper, we apply FL to CWS to leverage isolated data stored in different nodes and propose GCA to enhance the CWS model stored in the server. Specifically, our approach encodes the contextual information by associating the input characters with global character n-grams, and uses that information to guide the backbone model to make localized predictions. Experimental results under a simulated environment performed on five isolated nodes on CTB7 demonstrate the effectiveness of the proposed approach. Our approach outperforms the baseline model trained under the FL framework and achieves competitive results compared with the reference model that is trained on the union of the data from all nodes. Further analyses on identifying OOV justify the validity of the GCA mechanism to leverage the data on other nodes to facilitate localized prediction and demonstrate its great potential to be applied to real-world applications.

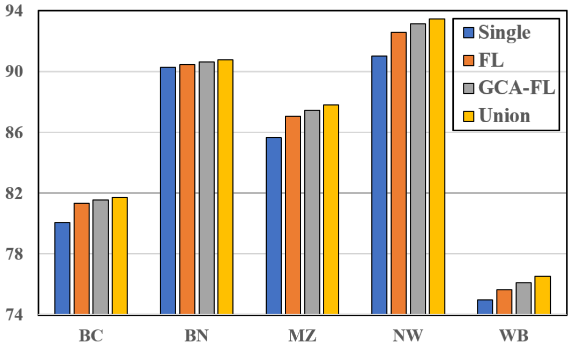

Figure 3: The recall of out-of-vocabulary words (OOV) of different BERT-based models (i.e., Single, FL, GCA-FL, and Union) on the test set of five nodes (genres).

Acknowledgements

This work is supported by Chinese Key-Area Research and Development Program of Guangdong Province (2020B0101350001) and NSFC under the project "The Essential Algorithms and Technologies for Standardized Analytics of Clinical Texts" (12026610). This work is also partially supported by Shenzhen Institute of Artificial Intelligence and Robotics for Society under the project "Automatic Knowledge Enhanced Natural Language Understanding and Its Applications" (AC01202101001).

References

Yoshinori Aono, Takuya Hayashi, Lihua Wang, Shiho Moriai, et al. 2017. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Transactions on Information Forensics and Security*, 13(5):1333–1345.

Guimin Chen, Yuanhe Tian, and Yan Song. 2020. Joint Aspect Extraction and Sentiment Analysis with Directional Graph Convolutional Networks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 272–279.

Guimin Chen, Yuanhe Tian, Yan Song, and Xiang Wan. 2021. Relation Extraction with Type-aware Map Memories of Word Dependencies. In *Findings of* *the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Mingqing Chen, Ananda Theertha Suresh, Rajiv Mathews, Adeline Wong, Cyril Allauzen, Françoise Beaufays, and Michael Riley. 2019. Federated Learning of N-Gram Language Models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 121–130.

Xinchi Chen, Zhan Shi, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial Multi-Criteria Learning for Chinese Word Segmentation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1193–1203.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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.

Jinlan Fu, Pengfei Liu, Qi Zhang, and Xuanjing Huang. 2020. RethinkCWS: Is Chinese Word Segmentation a Solved Task? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5676–5686, Online.

Suyu Ge, Fangzhao Wu, Chuhan Wu, Tao Qi, Yongfeng Huang, and Xing Xie. 2020. FedNER: Privacy-preserving medical named entity recognition with federated learning.

Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. 2018. Federated Learning for Mobile Keyboard Prediction. *arXiv preprint arXiv:1811.03604*.

Anxun He, Jianzong Wang, Zhangcheng Huang, and Jing Xiao. 2020. FedSmart: An Auto Updating Federated Learning Optimization Mechanism. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*, pages 716–724. Springer.

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.

Kaiyu Huang, Degen Huang, Zhuang Liu, and Fengran Mo. 2020a. A Joint Multiple Criteria Model in Transfer Learning for Cross-domain Chinese Word Segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3873–3882, Online.

Zhiqi Huang, Fenglin Liu, and Yuexian Zou. 2020b. Federated Learning for Spoken Language Understanding. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3467–3478, Barcelona, Spain (Online).

Zhen Ke, Liang Shi, Erli Meng, Bin Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Unified Multi-criteria Chinese Word Segmentation with BERT. *arXiv preprint arXiv:2004.05808*.

Jakub Konečný, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. 2016. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*.

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.

Dianbo Liu, Dmitriy Dligach, and Timothy Miller. 2019. Two-stage Federated Phenotyping and Patient Representation Learning. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 283–291, Florence, Italy.

Ji Ma, Kuzman Ganchev, and David Weiss. 2018. State-of-the-art Chinese Word Segmentation with Bi-LSTMs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4902–4908.

Mairgup Mansur, Wenzhe Pei, and Baobao Chang. 2013. Feature-based Neural Language Model and Chinese Word Segmentation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1271–1277, Nagoya, Japan.

Yuyang Nie, Yuanhe Tian, Yan Song, Xiang Ao, and Xiang Wan. 2020. Improving Named Entity Recognition with Attentive Ensemble of Syntactic Information. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4231–4245.

Wenzhe Pei, Tao Ge, and Baobao Chang. 2014. Max-margin Tensor Neural Network for Chinese Word Segmentation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 293–303.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar.

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.

Han Qin, Guimin Chen, Yuanhe Tian, and Yan Song. 2021. Improving Arabic Diacritization with Regularized Decoding and Adversarial Training. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Xipeng Qiu, Hengzhi Pei, Hang Yan, and Xuanjing Huang. 2019. Multi-Criteria Chinese Word Segmentation with Transformer. *arXiv preprint arXiv:1906.12035*.

Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. 2018. Multi-institutional Deep Learning Modeling without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation. In *International MICCAI Brainlesion Workshop*, pages 92–104.

Reza Shokri and Vitaly Shmatikov. 2015. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321.

Yan Song, Dongfeng Cai, Guiping Zhang, and Hai Zhao. 2009a. Approach to Chinese Word Segmentation Based on Character-word Joint Decoding. *Journal of Software*, 20(9):2236–2376.

Yan Song, Chunyu Kit, and Xiao Chen. 2009b. Transliteration of Name Entity via Improved Statistical Translation on Character Sequences. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*, pages 57–60, Suntec, Singapore.

Yan Song and Shuming Shi. 2018. Complementary Learning of Word Embeddings. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4368–4374.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2*, pages 175–180, New Orleans, Louisiana.

Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing Medical Conversations via Identifying Important Utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 717–729.

Yan Song and Fei Xia. 2012. Using a Goodness Measurement for Domain Adaptation: A Case Study on Chinese Word Segmentation. In *LREC*, pages 3853–3860.

Yan Song and Fei Xia. 2013. A Common Case of Jekyll and Hyde: The Synergistic Effect of Using Divided Source Training Data for Feature Augmentation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 623–631, Nagoya, Japan.

Yan Song, Tong Zhang, Yonggang Wang, and Kai-Fu Lee. 2021. ZEN 2.0: Continue Training and Adaption for N-gram Enhanced Text Encoders. *arXiv preprint arXiv:2105.01279*.

Dianbo Sui, Yubo Chen, Jun Zhao, Yantao Jia, Yuantao Xie, and Weijian Sun. 2020. FedED: Federated Learning via Ensemble Distillation for Medical Relation Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2118–2128, Online.

Weiwei Sun and Jia Xu. 2011. Enhancing Chinese Word Segmentation Using Unlabeled Data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 970–979.

Yuanhe Tian, Guimin Chen, and Yan Song. 2021a. Aspect-based Sentiment Analysis with Type-aware Graph Convolutional Networks and Layer Ensemble. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2910–2922, Online.

Yuanhe Tian, Guimin Chen, Yan Song, and Xiang Wan. 2021b. Dependency-driven Relation Extraction with Attentive Graph Convolutional Networks. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Yuanhe Tian, Wang Shen, Yan Song, Fei Xia, Min He, and Kenli Li. 2020a. Improving Biomedical Named Entity Recognition with Syntactic Information. *BMC Bioinformatics*, 21:1471–2105.

Yuanhe Tian, Yan Song, Xiang Ao, Fei Xia, Xiaojun Quan, Tong Zhang, and Yonggang Wang. 2020b. Joint Chinese Word Segmentation and Part-of-speech Tagging via Two-way Attentions of Auto-analyzed Knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8286–8296, Online.

Yuanhe Tian, Yan Song, and Fei Xia. 2020c. Supertagging Combinatory Categorial Grammar with Attentive Graph Convolutional Networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6037–6044.

Yuanhe Tian, Yan Song, Fei Xia, and Tong Zhang. 2020d. Improving Constituency Parsing with Span Attention. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1691–1703.

Yuanhe Tian, Yan Song, Fei Xia, Tong Zhang, and Yonggang Wang. 2020e. Improving Chinese Word Segmentation with Wordhood Memory Networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8274–8285, Online.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 168–171.

Yiou Wang, Jun'ichi Kazama, Yoshimasa Tsuruoka, Wenliang Chen, Yujie Zhang, and Kentaro Torisawa. 2011. Improving Chinese Word Segmentation and POS Tagging with Semi-supervised Methods Using Large Auto-Analyzed Data. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 309–317, Chiang Mai, Thailand.

Xinghua Zhu, Jianzong Wang, Zhenhou Hong, and Jing Xiao. 2020. Empirical Studies of Institutional Federated Learning for Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 625–634.