MSnet: A BERT-based Network for Gendered Pronoun Resolution

Zili Wang
CEIEC
Chengdu, China
wzlnot@gmail.com

Abstract

The pre-trained BERT model achieves a remarkable state of the art across a wide range of tasks in natural language processing. For solving the gender bias in gendered pronoun resolution task, I propose a novel neural network model based on the pre-trained BERT. This model is a type of mention score classifier and uses an attention mechanism with no parameters to compute the contextual representation of entity span, and a vector to represent the triple-wise semantic similarity among the pronoun and the entities. In stage 1 of the gendered pronoun resolution task, a variant of this model, trained in the fine-tuning approach, reduced the multi-class logarithmic loss to 0.3033 in the 5-fold cross-validation of training set and 0.2795 in testing set. Besides, this variant won the 2nd place with a score at 0.17289 in stage 2 of the task.

The code in this paper is available at: https://github.com/ziliwang/MSnet-for-Gendered-Pronoun-Resolution

1 Introduction

Coreference resolution is an essential field of natural language processing (Sukthanker et al., 2018) and has been widely used in many systems such as dialog system (Niraula et al., 2014; Wessel et al., 2017), relation extraction (Wang et al., 2018) and question answer (Vicedo and Ferrández, 2000). Up to now, various models for coreference resolution have been proposed, and they can be generally categorized as (1) mention-pair classifier model (Webster and Nothman, 2016), (2) entity-centric model (Clark and Manning, 2015), (3) ranking model (Lee et al., 2017, 2018). However, some of these models implicate gender bias (Koolen and van Cranenburgh, 2017; Rudinger et al., 2018). To address this, Webster et al. (2018) presented and released Gendered Ambiguous Pronouns (GAP) dataset.

Recent work indicated that the pre-trained language representation models benefit to the coreference resolution (Lee et al., 2018). In the past years, the development of deep learning methods of language representation was swift, and the newer methods were shown to have significant effects on improving other natural language processing tasks(Peters et al., 2018; Radford and Salimans, 2018; Devlin et al., 2018). The latest one is Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), which is the cornerstone of the state of the art models in many tasks.

In this paper, I present a novel neural network model based on the pre-trained BERT for the gendered pronoun resolution task. The model is a kind of mention score classifier, and it is named as Mention Score Network (MSNet in short) and trained on the public GAP dataset. In particular, the model adopts an attention mechanism to compute the contextual representation of the entity span, and a vector to represent the triple-wise semantic similarity among the pronoun and the entities. Since the MSnet can not be tuned in a general way, I employ a two-step strategy to achieve the tuning-fine, which tunes the MSnet with freezing BERT firstly and then tunes them together. Two variants of MSnet are submitted in the gendered pronoun resolution task, and their logarithmic loss of local 5-fold cross-validation of train dataset is 0.3033 and 0.3042 respectively. Moreover, in stage 2 of the task, they acquired the score at 0.17289 and 0.18361 respectively, by averaging the predictions on the test dataset, and won the 2nd place in the task.

Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing, pages 89–95
Florence, Italy, August 2, 2019.

2 Model

As the target of the Gendered Pronoun Resolution task is to label the pronoun with whether it refers to entity A, entity B, or NEITHER. I aim to learn the reference probability distribution $P(E_i|D)$ from the input document D:

$$P(E_i|D) = \frac{\exp(s(E_i|D))}{\sum_{j \in E} \exp(s(E_j|D))}$$

where E_i is the candidate reference entity of pronoun, $E = \{\text{A}, \text{B}, \text{NEITHER}\}$ and s is the score function which is implemented by a neural network architecture, which is described in detail in the following subsection.

2.1 The Mention Score Network

The mention score network is build on the pre-trained BERT model (Figure 1). It has three layers, the span representation layer, the similarity layer, and the mention score layer. They are described in detail in the following part.

Span Representation Layer: The contextual representation is crucial to accurately predict the relation between the pronouns and the entities. Inspired by Lee et al. (2017), I adopt the hidden states of transformers of the pre-trained BERT as the contextual representation. As Devlin et al. (2018) showed that the performance of the concatenation of token representations from the top hidden layers of pre-trained Transformer of BERT is close to fine-tuning the entire model, the top hidden states will be given priority to compute the representation of entity spans. Since most entity spans consist of various tokens, the contextual representation of them should be re-computed to maintain the correspondence. I present two methods to re-compute the span representations: 1) **Meanpooling method**:

$$x^*_{(j,l)} = \frac{1}{\hat{N}} \sum_{i \in \text{Span}_j} x_{(i,l)}$$

where $x_{(i,l)}$ denotes the hidden states of i-th token in l-th layer of BERT, and $x^*_{(j,l)}$ denotes the contextual representation of entity span j, and $\hat{N}$ is the token counts of span j. 2) **Attention mechanism:** Instead of weighting each token equality, I adopt the attention mechanism to weight the tokens by:

$$s_{(i,l)} = \frac{1}{\sqrt{d_H}} \text{norm}(x_{(i,l)}) \cdot x_{(p,l)}$$

$$a_{(i,j,l)} = \frac{\exp(s_{(i,l)})}{\sum_{k \in \text{Span}_j} \exp(s_{(k,l)})}, i \in \text{Span}_j$$

$$x^*_{(j,l)} = \sum_{i \in \text{Span}_j} a_{(i,j,l)} x_{(i,l)}$$

The weights $a_{(i,j,l)}$ are learned automatically from the contextual similarity $s_{(i,l)}$ between pronoun $x_{(p,l)}$ and the token $x_{(i,l)}$ in the span j. Different from the commonly used attention functions, the above one has no parameters and is more space-efficient in practice. The scaling factor d_H denotes the hidden size of BERT and is designed to counteract the effect of extremely small gradients caused by the large magnitude of dot products (Vaswani et al., 2017).

Similarity Layer: Inspired by the pairwise similarity of Lee et al. (2017), I assume a vector $\hat{\mathbf{s}}_l$ to represent the triple-wise semantic similarity among the pronoun and the entities of l-layer in BERT:

$$a_l = x^*_{(a,l)}$$

$$b_l = x^*_{(b,l)}$$

$$p_l = x_{(p,l)}$$

$$\hat{\mathbf{s}}_l = \mathbf{W}^T[p_l, a_l, b_l, a_l \circ p_l, b_l \circ p_l] + \mathbf{b}$$

where a_l, b_l and p_l denote the contextual representation of the pronoun, entity A and entity B of the l-th layer in BERT, $\cdot$ denotes the dot product and $\circ$ denotes the element-wise multiplication. The $\hat{\mathbf{s}}_l$ can be learned by a single layer feed-forward neural network with the weights $\mathbf{W}$ and the bias $\mathbf{b}$.

Mention Score Layer: Mention score layer is also a feed-forward neural network architecture and computes the mention scores given the distance vector $\mathbf{d}$ between the pronoun and its candidate entities and the concatenated similarity vector $\hat{\mathbf{s}}$:

$$d_a = \tanh(w_{\text{dist}}(\text{START(A)}-\text{START(P)}) + b_{\text{dist}})$$

$$d_b = \tanh(w_{\text{dist}}(\text{START(B)}-\text{START(P)}) + b_{\text{dist}})$$

$$\mathbf{d} = [d_a, d_b]$$

$$\hat{\mathbf{s}} = [\hat{\mathbf{s}}_0, \hat{\mathbf{s}}_1, ..., \hat{\mathbf{s}}_l, ..., \hat{\mathbf{s}}_L]$$

$$s(E_i|D) = \mathbf{W}_{E_i} \cdot [\hat{\mathbf{s}}, \mathbf{d}] + b_{E_i}$$

where d_a (or d_b) denotes the distance encoding of entity A (or B), $\hat{\mathbf{s}}_l$ denotes the similarity vector computed by the representation of the l-th layer in BERT. L is the total layers for representation, and START denotes the index of the start token of

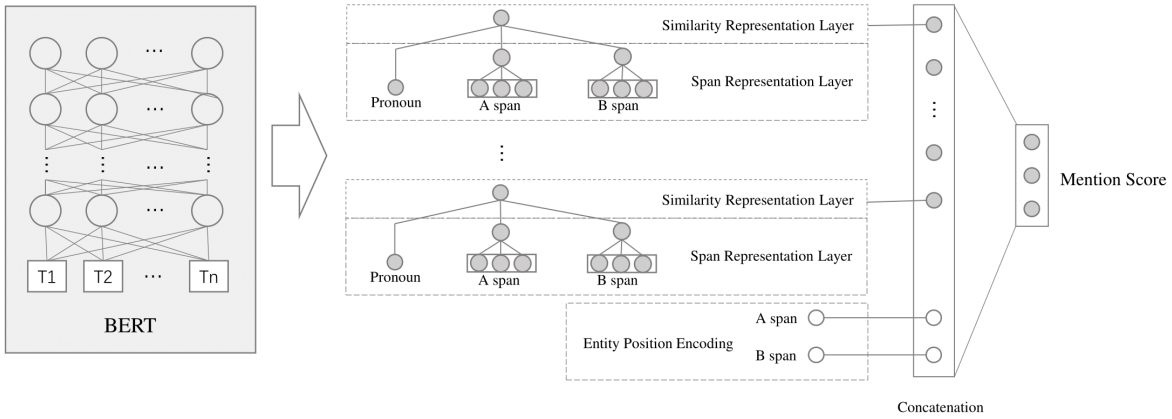

Figure 1: The architecture of MSnet.

the span. w_{dist} is a learnable weight for encoding the distance which corresponds to a learnable bias b_{dist} and $\mathbf{W}_{E_i}$ is the learnable weights for scoring entity E_i which corresponds to a learnable bias b_{E_i}.

3 Experiments

I train the model on the Kaggle platform by using scripts kernel which using the computational environment from the docker-python [1]. I employ pytorch as the deep learning framework, and the `pytorch-pretrained-BERT` package [2] to load and tune the pre-trained BERT model.

3.1 Dataset

The GAP Coreference Dataset [3] (Webster et al., 2018) has 4454 records and officially split into three parts: development set (2000 records), test set (2000 records), and validation set (454 records). Conforming to the stage 1 of Gendered Pronoun Resolution [4] task, the official test set and validation set are combined as the training dataset in the experiments, while the official development set is used as the test set correspondingly.

[1] `https://github.com/Kaggle/docker-python`

[2] `https://github.com/huggingface/pytorch-pretrained-BERT`

[3] `https://github.com/google-research-datasets/gap-coreference`

[4] `https://www.kaggle.com/c/gendered-pronoun-resolution`

3.2 Preprocessing

In the experiments, the WordPiece is used to tokenize the documents. To ensure the token counts less than 300 after tokenizing, I remove the head or tail tokens in a few documents. Next, the special tokens `[CLS]` and `[SEP]` are added into the head and end of the tokens sequences.

3.3 Hyper-parameters

Pre-trained BERT model: As increasing model sizes of BERT may lead to significant improvements on very small scale tasks (Devlin et al., 2018), I explore the effect of $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$ in the experiments. I employ the `uncased_L-12_H-768_A-12` [5] as the $\text{BERT}_{\text{BASE}}$ and `cased_L-24_H-1024_A-16` [6] as the $\text{BERT}_{\text{LARGE}}$, and both of them are transformed into the pytorch-supported format by the script in `pytorch-pretrained-BERT`.

Hidden Layers for Representation: Devlin et al. (2018) showed that using the representation from appropriate hidden layers of BERT can improve the model performance, the hidden layers L (described in Section 2) is therefore utilized as a hyper-parameter tuned in the experiments.

Dimension of Similarity Vector: Since a vector is used to represent the task-specific semantic similarity, its dimension $\hat{\mathbf{s}}_{dim}$ may have potential influence the performance. A smaller dimension

[5] `https://storage.googleapis.com/bert_models/2018_10_18/uncased_L-12_H-768_A-12.zip`

[6] `https://storage.googleapis.com/bert_models/2018_10_18/cased_L-24_H-1024_A-16.zip`

will partly lose information, while a bigger one will cause generalization problems.

Span Contextual Representation: As section 2 described, both the meanpooling and attention method can be used to compute the contextual representation of the tokens span of the entity. Therefore, the choice of them is a hyper-parameter in the experiment.

Tunable Layers: I use two different approaches to train the MSnet model. The first one is the feature-based approach which trains MSnet with freezing the BERT part. The second one is the fine-tuning approach, which tunes the parameters of BERT and MSnet simultaneously. Howard and Ruder (2018) showed the discriminative fine-tuning gets a better performance than the ordinary, which possibly means that the pre-trained language model has a hierarchical structure. One possible explanation is that the lower hidden layers extract the word meanings and grammatical structures and the higher layers process them into higher-level semantic information. In this, I freeze the embedding layer and bottom hidden layers of BERT to keep the completeness of word meaning and grammatical structure and tune the top hidden layers L_{tuning}.

3.4 Training Details

For improving the generalization ability of the model, I employ the dropout mechanism (Srivastava et al., 2014) on the input of the feed-forward neural network in the similarity layer and the concatenation in the mention score layer. The rate of dropout is set at 0.6 which is the best setting after tuned on it. I also apply the dropout on the representation of tokens when using the attention mechanism to compute the contextual representation of span, and its dropout rate is set at 0.4. Additionally, I adopt the batch normalization (Ioffe and Szegedy, 2015) before the dropout operation in the mention score layer. As introduced in section 3.3, I use the feature-based approach and the fine-tuning approach separately to train the MSnet, and the training details are described in the following.

Feature-based Approach: In the feature-based approach, I train the model by minimizing the cross-entropy loss with Adam (Kingma and Ba, 2014) optimizer with a batch size of 32. To adapt to the training data in the experiments, I tuned the learning rate and found a learning rate of 3e-4 was the best setting. The maximum epoch set at 30 and early stopping method is used to prevent the over-fitting of MSnet.

Fine-tuning Approach: In the fine-tuning approach, the generic training method was not working. I adopt a two-step tuning strategy to achieve the fine-tuning. In step 1, I train the MSnet in the feature-based approach. And in step 2, MSnet and BERT are tuned simultaneously with a small learning rate.

Since the two steps have the same optimization landscape, in step 2, the model may not escape the local minimum where it entered in step 1. I adopt two strategies of training in step 1 to reduce the probability of those situations: 1) premature. The MSnet is trained to under-fitting by using a small maximum training epoch which is set at 10 in the experiments. 2) mature. In this strategy, MSnet is trained to proper-fitting, and it is applied by adopting a weight decay at 0.01 rate, an early stopping at 4 epoch, and the maximum training epoch at 20 in the experiments. In addition, other training parameters of the two strategies have the same setting as in the feature-based approach.

In step 2, I also trained the model by minimizing the cross-entropy loss but with two different optimizers. For BERT, I used the Adam optimizer with the weight decay fix which implemented by `pytorch-pretrained-BERT`. For MSnet, the generic Adam was used. Both of the two optimizers are set with a learning rate at 5e-6 and a weight decay at 0.01. The maximum training epoch is set at 20, and the early stopping is set at 4 epoch. The batch size was 5 as the GPU memory limitation.

3.5 Evaluation

I report the multi-class logarithmic loss of the 5-fold cross-validation on train and the average of their predictions on the test. Also, the running time of the scripts is reported as a reference of the performance of the MSnet.

4 Results and Discussion

4.1 Feature-based Approach

The results of MSnet variants trained in feature-based approach are shown in Table 1. The comparison between model #1 and model #2 shows that the combination of the top 4 hidden layers for contextual representation is better than the top

Model#	BERT	L	$\hat{\mathrm{s}}_{\mathrm{dim}}$	Span	5-fold CV on train	test	runtime(s)
1	BASE	1	32	Meanpooling	0.5247±0.0379	0.4891	232.8
2	BASE	4	32	Meanpooling	0.4699±0.0431	0.4270	317.3
3	LARGE	4	32	Meanpooling	0.4041±0.0532	0.3819	358.3
4	LARGE	8	32	Meanpooling	0.3783±0.0468	0.3519	372.2
5	LARGE	12	32	Meanpooling	0.3879±0.0461	0.3546	415.4
6	LARGE	8	8	Meanpooling	0.3758±0.0430	0.3490	436.2
7	LARGE	8	16	Meanpooling	0.3736±0.0465	0.3488	415.0
8	LARGE	8	64	Meanpooling	0.3780±0.0441	0.3518	447.6
9	LARGE	8	16	Attention	0.3582±0.0435	0.3349	828.2

Table 1: Results of Feature-based Aproach.

Model#	Based Model	method	L_{tuning}	5-fold CV on train	test	runtime(s)
10	#9	premature	12	0.3033±0.0367	0.2795	6909.5
11	#9	mature	12	0.3042±0.0352	0.2856	7627.7
12	#9	mature	8	0.3110±0.0352	0.2876	8928.1
13	#9	mature	16	0.3185±0.0465	0.2820	7763.4
14	#9	mature	24	0.3169±0.0440	0.2843	8695.4

Table 2: Results of Fine-tuning Aproach.

layer. The possible reason is that the semantic information about gender may be partly transformed to the higher level semantic information during the hidden layers in BERT. In addition, changing $\mathrm{BERT}_{\mathrm{BASE}}$ to the $\mathrm{BERT}_{\mathrm{LARGE}}$ reduces the loss in 5-fold CV on train from 0.4699±0.0431 to 0.4041±0.0532, which demonstrate increasing model size of BERT can lead to remarkable improvement on the small scale task. The exploration of contextual representation layers shows the proper representation layers is proportionate to the number of hidden layers of BERT. In other words, the modeling ability of $\mathrm{BERT}_{\mathrm{LARGE}}$ is more powerful than $\mathrm{BERT}_{\mathrm{BASE}}$ by using a more complex function to do the same work.

The comparison among the model #4, model #6, model #7 and model #8 shows the dimension of the similarity vector has a slight affection for the performance of MSnet (Table 1) and the best loss is 0.3736±0.0465 with the dimension set at 16. Changing the method for computing the span contextual representation from meanpooling to attention mechanism reduces the loss in CV on train by ∼0.02, which demonstrates that the attention mechanism used in the experiment is effective to compute the contextual representation of the entity span. To the best of my knowledge, it is a novel attention mechanism with no learnable parameters and more space-efficient and more explainable in practice.

4.2 Fine-tuning Approach

The experiments in fine-tuning approach was based on model #9, and the results are shown in table 2. The comparison between model #10 and model #11 shows that their difference on performance is slight. Also, both of them are effective to the fine-tuning of MSnet and reduce loss in the CV of train by ∼0.054 compared to the feature-based approach. Furthermore, the tuning on L_{tuning} shows the best setting is tuning top 12 hidden layers in BERT, and more or fewer layers will reduce the performance of MSnet. The possible reason is that tuning fewer layers will limit the ability of the transformation from basic semantic to gender-related semantic while tuning more bottom layers will damage the extraction of the underlying semantics when training on a small data set.

As the apporach transformed from the feature-based to the fine-tuning, the intentions of some hyper-parameters were changed. The obvious one is the hidden layers for contextual representation, which is used to combine the semantic in each hidden layers in the feature-based approach and changed to constrain the contextual representation to include the same semantic in fine-tuning approach. Although, the change on the intentions was not deliberate, the improvement on the per-

formance of the model was observed in the experiments.

4.3 Results in Stage 2

The gendered pronoun resolution was a two-stage task, and I submitted the model #10 and #11 in stage 2 as their best performances in 5-fold cross-validation of the training dataset. The final scores of the models were 0.17289 (model #10) and 0.18361 (model #11). This result featurely demonstrates the premature strategy is better than the mature one and can be explained as former one keeps more explorable optimization landscape in step 2 in the fine-tuning approach.

5 Conclusion

This paper presented a novel pre-trained BERT based network model for the gendered pronoun resolution task. This model is a kind of mention score classifier and uses an attention mechanism to compucate the contextual representation of entity span and a vector to represent the triple-wise semantic similarity among the pronoun and the entities. I trained the model in the feature-based and the two-step fine-tuning approach respectively. On the GAP dateset, the model trained by the fine-tuning approach with premature strategy obtains remarkable multi-class logarithmic loss on the local 5-fold cross-valication at 0.3033, and 0.17289 on the test dataset in stage 2 of the task. I believe the MSnet can serve as a new strong baseline for gendered pronoun resolution task as well as the coreference resolution. The code for training model are available at: `https://github.com/ziliwang/MSnet-for-Gendered-Pronoun-Resolution`

Acknowledgments

I want to thank the kaggle company for its public computing resources.

References

Kevin Clark and Christopher D Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1405–1415.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Corina Koolen and Andreas van Cranenburgh. 2017. These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. *arXiv preprint arXiv:1804.05392*.

Nobal B Niraula, Vasile Rus, Rajendra Banjade, Dan Stefanescu, William Baggett, and Brent Morgan. 2014. The dare corpus: A resource for anaphora resolution in dialogue based intelligent tutoring systems. In *LREC*, pages 3199–3203.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Alec Radford and Tim Salimans. 2018. GPT: Improving Language Understanding by Generative Pre-Training. *arXiv*, pages 1–12.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2018. Anaphora and Coreference Resolution: A Review.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

José L Vicedo and Antonio Ferrández. 2000. Importance of pronominal anaphora resolution in question answering systems. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. 2018. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49.

Kellie Webster and Joel Nothman. 2016. Using mention accessibility to improve coreference resolution. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 432–437.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. In *Transactions of the ACL*, page to appear.

Michael Wessel, Girish Acharya, James Carpenter, and Min Yin. 2017. An ontology-based dialogue management system for virtual personal assistants. In *Proceedings of the 8th International Workshop On Spoken Dialogue Systems (IWSDS)*.